Journal of Theoretical and Applied Information Technology
31st July 2023. Vol.101. No 14
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MACHINE LEARNING PIPELINE APPROACH TO SECURE IOT-BASED SMART CITIES

ABDESSAMAD BADOUCH[1] , SALAHEDDINE KRIT[2]

Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Agadir, Morocco
E-mail: [1] abdessamad.badouch@gmail.com, [2] salahddine.krit@gmail.com

ABSTRACT

The data stored and transmitted in IoT-based smart cities is often huge and highly sensitive. One of the major threats to data in IoT-based smart cities is network intrusions and attacks. Various algorithms have been proposed to mitigate these attacks with high accuracy. However, most of them have major flaws, such as high detection times and the limitation of only being able to perform binary classification of network traffic. In this paper, we propose a new Machine Learning (ML)-based approach for intrusion detection in IoT-based smart cities. The proposed approach performs the detection in a pipeline of three phases: (1) preprocessing of incoming network traffic; (2) binary classification of traffic data into normal or attack; and (3) perform a final multi class classification phase. The experiments were performed on the TON-IoT public dataset, and after a comparison study of several ML algorithms, we settled to use Decision Tree (DT) for the binary classification phase and Random Forest (RF) for multi class classification. In addition to high accuracy, our approach achieves better computation time, and more fine-grained detection of attacks, which are highly important for intrusion detection systems. This aspect should not be overlooked in the context of IOT devices, which usually have very limited computational resources.

Keywords: *IOT, Smart Cities, Machine Learning, Security, Privacy*

1. INTRODUCTION

Owing to better facilities and standard of life, a significant amount of the human population has moved from rural to urban areas in the past few decades. As per United Nations the 54% of the global population used to live in urban areas in 2014, however, by 2050 it is estimated that roughly 68% of the global population will be residing in the cities [1]. It is also estimated that by 2030, there will be more than 650 cities, out of which almost 40 will be megacities. Moreover, most of these cities will be in the regions that have potential of significant economic growth [2]. Such an exponential increase in urban population will be challenging for governments from several perspectives. The infrastructure of education, healthcare, energy, transportation and housing must be upgraded, moreover, the policies regarding management of city, environment and security must be revised [1]. Various strategies are employed by the modern cities that help in their socio-economic progress, sustainability, and environment. The cities will be transformed into smart cities that will not only be able to house the increased influx of population but will also offer a better standard of living. Number of smart city projects are currently going on around the world [3].

An important feature that will help transform a conventional city to a smart is the effective and widespread use of artificial intelligence (AI). Over the past few years, the appeal of AI has increased in the smart city research. Move rover, number of countries have started to use AI for achieving the United Nation’s sustainable development goals (SDG) [4]. The use of AI in any smart city can offer several advantages. These include, but not limited to, efficient use of energy and water resources, better management of waste disposal, monitoring of environmental impact and supervision of traffic congestion and noise pollution. Moreover, the AI can also help in gathering and analysis of the city data. This can yield new information regarding trends and dynamics of the city [5]. Using AI, a number of smart city applications have been designed, that will considerably improve the standard of livings, reduce environmental impact, sustainability of city and effective use of its resources [6]. Courtesy this, it is imperative for governments to regularly collect diverse data from cities and take steps to improve the quality of life for smart city residents.

2. ARTIFICIAL INTELLIGENCE FOR THE SECURITY OF SMART CITIES

As we can see, one of the vital aspects of the smart city is its ability to collect data and then transmit it for further analysis. In order to effectively connect various devices with the internet in context of smart city, the technologies like Information and Communication Technology (ICT) and Internet of things (IOT) can be extremely helpful. The IOT is a state-of-the-art technology that is used to facilitate communication over the internet between different heterogenous devices. Such devices may include, but not limited to sensors, routers and electronic devices. One of the revolutionary examples of IOT is the integration of home-based patient monitoring device with the remotely located physician. Such a system not only provides fast, accurate, and cost-effective diagnosis, it can also keep track and history of the patient [7]. Since sensors of different natures are connected to the internet in IOT, the threat of cybersecurity is real. This threat is grave also since different devices are often using different protocols to communicate with one another, making a uniform safety strategy challenging. Moreover, many intrusions detection devices consume lot of energy, resulting in their limited use [8].

The strategies employed by hackers to intrude in the smart city IOT system are improving and hence the need to protect the system against them as well. Most of the devices connected to IOT work autonomously and are often connected to cellular networks. This makes their continuous monitoring impractical against any eavesdropping [9]. Furthermore, due to limited processing capacity and energy requirements, the addition of intrusion detection device with IOT devices is also not feasible. Here again, the benefits of AI can be reaped, as it can be used to log and analyze network behavior. Any deviation of the network parameters from the normal behavior can be a signal of the intrusion attempt. In addition to being an effective strategy, such software-based approaches are extremely cost effective as well. An AI-based management solution, to be setup near the IOT edge and big data collection points, for securing the IOT of the smart city is proposed in this paper [10]. An intrusion detection system based on the machine learning is suggested by Rahman et al. in [11]. For the detection of malicious URL, an approach based on deep belief network (DBN) and deep neural network (DNN) has been investigated by Selvaganapathy et al. in [12]. For the real-time detection of cyber threat, a method based on support vector machine (SVM) and DBN has been proposed in [13]. A hybrid deep learning (DL) and convolutional neural network (CNN) method has also been investigated for the classification between fake and true news [14]. Furthermore, to protect smart cities against the cyber threat a system, using IOT-based radio frequency, for the detection of attack points has been investigated in this paper [15]. These strategies for data protection and for the protection against cyber-attacks are crucial for smart cities as the protection of data is in fact the protection of the citizen of the smart city. Furthermore, the machine learning-based system has proven to be quite effective against cyber-attack and can strengthen the network. Lastly, machine learning can be useful in pattern recognition, managing and storing public data, analysis of network behavior and detection of security threats [16].

3. SECURITY AND PRIVACY THREATS IN IOT-BASED SMART CITY

The IOT is characterized by the fact that diverse devices and sensors are connected to the internet. Most of these devices are purpose-built and have limited computational power. Furthermore, all these devices have different vendors, and they use separate communication protocols and threads. Moreover, these devices lack a standard operating system, and they often work with one another over insecure wireless media. All these limitations make it difficult to develop a standard security protocol that can work for all the devices [17]. Cybersecurity research in recent years has yielded several software-based techniques that can be rolled on across the devices in IOT. Moreover, such software needs to be continuously updated to be effective against new and emerging security threats [18].

Like IOT devices, the security threats to this system are also varied. Furthermore, the IOT network is at risk from security threats both from inside and outside of the network [19]. Several such threats have been identified by the International Engineering Task Force (IETF). These threats include Denial of Service attack (DoS), Man in The Middle attack (MiTM), privacy risks, eavesdropping and replacement of firmware with malicious code [20]. Different security perils compromise availability, integrity, and privacy of the network. Hence, security threats comprise availability and integrity of network, while privacy attacks comprise the confidentiality of the network. The figure 1. Illustrates the different types of security and privacy risks faced by IOT devices.

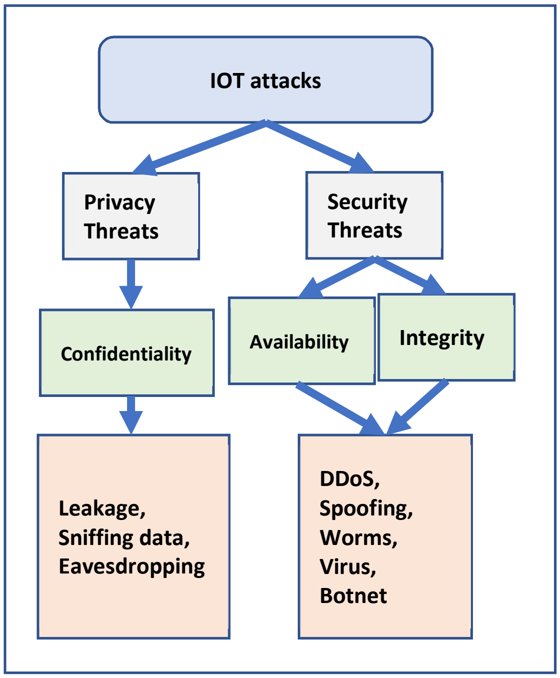

Figure 1: Different types of security and privacy threats being faced by IOT devices. Both availability and integrity are part of security

3.1 Security threats

The most common security threat is the denial-of-service attack (DOS). From the hacker's perspective, this attack is easy to execute due to inadequate security features in IOT devices. The aim of denial-of-service attack is to make the system unavailable for users. It is commonly executed by sending huge number of unauthorized requests that will eventually choke the resources of the system. A sub-class of DOS is known as distributed DOS (DDOS), where multiple IOT devices are targeted to overwhelm the network with bogus traffic [21]. Other kinds of denial-of-service attacks exist as well, namely, phishing, network probing, information stealing etc., however, umbrella term used for all such attacks is Botnet attack.

Another, and somewhat old, type of security threat is the MiTM attack. As the name suggests, in this kind of threat an attacker gets access to the data being communicated between two parties within the network or, even worse, he is able to change the data being communicated. At times, malicious software is introduced in the system to cause disruptions. Such software are commonly known as malwares and they can then further be classified as worms, spyware, trojan horses etc. [22]. The health care sector and the smart traffic management systems are vulnerable to malware attacks.

3.2 Privacy threats

Privacy threats typically concern data privacy. Smart cities involve large amounts of data collected from various sensors and from citizens. Such data is helpful in estimating the current and future needs of the city. The threats involving data are data theft, data substitution and impersonation [23]. Such threats, if not, addressed can lead to unpleasant consequences. The MiTM attack can also be used for invasion of privacy, where the hacker simply listens to the communication that he is interested in. IOT devices, such as smart phones and smart watches are prone to eavesdropping and sniffing. A hacker can also impersonate someone that he is not and make the user send him information that he was not authorized to have.

An effective mechanism is crucial for the smart system that can safeguard the IOT devices against all these cyber threats. Different aspects of threats that IOT devices are prone to, are highlighted in figure 2. The users, and most importantly, the management of the smart city must be knowledgeable of these different types of threats to come up with the right strategy. AI can be extremely beneficial for devising mechanisms that make the network safe against security and privacy threats. Different machine learning and deep learning base strategies exist to boast the security and privacy of the IOT in the realm of smart cities. In the next section, we will describe some of those techniques.

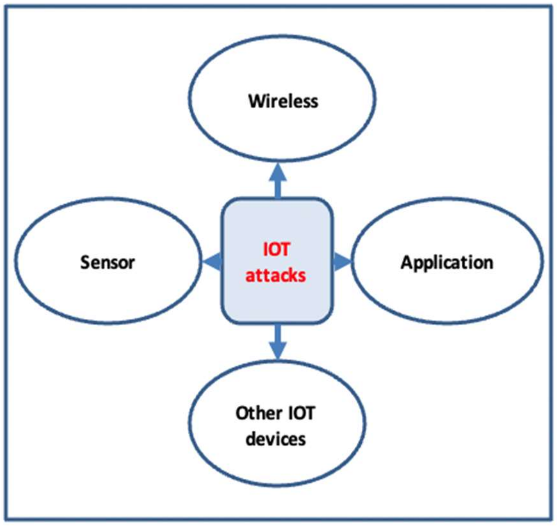

Figure 2: Different aspects of attack on IOT. Both hardware sensor and software are prone to cyber threat.

4. MACHINE LEARNING TECHNIQUES FOR SECURITY IN SMART CITIES

Machine learning approaches have successfully been used in economics and other fields [24]. The approaches work on the data set and can predict and estimate the abnormal behavior with minimal human intervention. They often needed to be trained first. For example, the data related to network traffic under normal conditions can be used to train the model. Once it is done, the model will be smart enough to predict the intrusion or any divergence from the normal behavior and hence can automatically and autonomously identify the cyber threat.

4.1 Decision tree

The decision tree (DT) algorithm is an example of supervised machine learning, that can be used for classification and regression problems. Regression can be termed as prediction or estimation; however, classification can be used for categorization of entities into distinct classes. To be precise the DT algorithm works by splitting the data iteratively using a tree-based structure in order to maximize the measure of purity [25].

The random forest (RF) regressor is constituted using multiple DTs, where each individual tree is trained on the segment of training data to minimize the error between the prediction and the ground truth. In the perspective of the smart city, the DT has been used to estimate the characterization of traffic and its performance has been compared with other machine learning algorithms. As per literature, the accuracy metric for DT was 99.18% [26]. In another study, the DT was used for the prediction of pandemic and in comparison, with other machine learning algorithms it gave the 99% accuracy for estimation [27].

4.2 Support Vector Machine

The support vector machine (SVM) algorithm is one of the widely used machine learning algorithms, that can work as a regressor and a classifier using a learned model. The SVM works by considering the training data as set of points in feature space. The algorithm then finds the hyperplanes that best classify data into different classes based on the features. Using blockchain based encrypted data gathered from IOT, Shen et al. devised an SVM algorithm for security and privacy in IOT smart city [28]. Their results report that the algorithm provided sufficient security and confidentiality of the data in the specific task. In another study, the SVM was employed to recognize a cyberattack, its performance was compared with other machine learning algorithms, however, the performance of SVM was not found to be very promising [29]. Using Kyoto 2006+ data set, the authors of another study employed six machine algorithms to detect the anomalies and opposite decision in computer network in context of smart city. Their results show that all the algorithms gave promising accuracy, however, the accuracy of SVM was not the best among other techniques [30].

4.3 Artificial Neural Network

The artificial neural network (ANN) is the simple way of designing an intelligent system that is capable of self-learning, inspired by the neurons present in the human brain. Such system is smart enough to extract knowledge from the training data, without the need to be programmed about task-specific features [36]. Hence, the core idea behind any ANN is to execute a task without any a-priori knowledge about the data, subsequently, the ANN can also extract the discriminative features form the input data. Owing to their versatility, the ANN have seen their applications in forecasting, cure-fitting and regression. Another aspect of the ANN, that makes them desirable is that they can be used on very large data sets, all the while being simple and cost effective.

The ANN has been used to ward off cyberattacks successfully. In one study, the classification of unknown data sample into UNSW-NB15 and CICIDS20017 datasets. The accuracy of ANN classifier was found to be 80.69% and 78.23% respectively [29]. Using dynamic neural network, another study developed a learning model that can predict the performance of IOT communication system in context of smart city. Moreover, the results reported high accuracy and elimination of errors [31]. Another study compared various machine learning algorithms to detect the intrusion in the cloud traffic network of smart cities. As per their published results the neural network based approach reached the accuracy of 99.33% [32]. Using a combination of convolutional neural network (CNN) and quasi-recurrent neural network (QNN), a study developed a model that can detect cyber threats. After being tested on two datasets, namely BoT-IOT and TON-IOT, the developed model achieved accuracy of 99.99% for both data sets [33].

5. HYBRID AND DEEP LEARNING TECHNIQUES FOR SECURITY IN SMART CITIES

This section illustrated the ensemble and deep learning-based approaches that can be used for security and privacy in smart cities.

5.1 Ensemble and Hybrid technique

The hybrid approaches combine two or more machine learning algorithms with the aim of exploiting their advantages for executing the task at hand. An example being the integration of extreme learning machine (ELM) and the SVM for prediction and to improve optimization and chemical industry in this paper [34]. As per their results the combination of ELM and SVM gave better accuracy in comparison to when SVM and ANN were used individually. Ensemble approaches, on the contrary, combine multiple machine learning approaches but with different goals, such decreasing bias or increasing performance. The underlying idea behind ensemble approaches is that using multiples for executing the same problem will produce a model with better performance. How exactly different models will be combined is known as meta-algorithm. Bagging is one of the meta-algorithms that trains the ML algorithms in parallel and gives output as a deterministic average of their individual outputs. Bagging has often been used to improve the performance of DTs used in random forest [35]. Boosting is another meta-algorithm where individual ML algorithms (also called as weak learners in this context) are trained sequentially in adaptive fashion and their output is combined deterministically. Moreover, advanced combination techniques can be used to improve both decision and outputs of the base models [36].

5.2 Deep Learning Techniques

Among the advanced ML algorithms, the deep learning (DL) algorithm has shown great promise in the last few years. DL is a subclass of artificial intelligence, that can work on unstructured data as well. The DL employes multilayer neural network that is modeled on the human brain. Hence, it can find the suitable differentiating features in the data all by itself. The term ‘deep’ refers to the number of transformational steps that will create those features. Some of the commonly used DL algorithms CNN, auto encoders (AEs) and recurrent neural network (RNN) [37]. The DL techniques have been used in several domains, including audio and speech processing, pattern recognition, traffic analysis and big data management.

The convolutional neural network (CNN) is one of the most widely used variants of DL, which is inspired from the visual system of living beings. Generally, CNN is composed of four components. Those components are convolutional layer, pooling layer, activation function and fully connected layer [38]. The convolutional layer is the main block of the CNN, and it consists of a set of filters, who are supposed to be taught during the whole training cycle. Each filter uses convolution operation to produce the activation map. The pooling layer optimizes the number of parameters required by the network and produces the pooled feature map. The activation function improves the nonlinear abilities of the network and improves the classification accuracy [39]. The fully connected layer finally produces the appropriate output, based on the desired task.

The recurrent neural network (RNN) is a type of DL algorithm based on connections, that recurrently capture the changes in sequences in a network consisting of different nodes. The RNN have been used to detect malicious attack on smart grids in the smart cities and have reported to achieve the accuracy of more than that of 90% [40]. The long-short term memory (LSTM) is the most widely used variant of RNN. It consists of so-called cells which can store and mimic the temporal behavior with long-term dependencies. The LSTM model consists of three gates. These are input, output and forget. The function of these gats is to regulate the flow of information and to take decisions regarding which information is relevant and should be kept and which information is not relevant and can be forgotten [41].

The DL techniques have great appeal owing to their applicability. This makes them a promising tool in comparison to ML techniques due to their better performance. As the number of smart cities grows, so does the usage of DL techniques in various IOT applications as well.

6. RELATED WORKS

Number of studies of applied simple and advanced machine learning algorithms on diverse datasets. In this section, we will briefly review the studies that have applied decision tree and random forest classifier. In one study, the decision tree classifier was used to classify the network traffic. As per results the decision tree provided the best classification with accuracy of 99% [26]. The decision tree algorithm was also used for the pandemic prediction in smart city, and it gave accuracy of 99.2% percent and performed better than

KNN and LR [43]. In this paper, the decision tree algorithm was used for the prediction of energy stress due to climate change in the smart city, as per results almost 80% lower cooling stress can be achieved [44]. In yet another study, the decision tree in combination of DBN was used to detect the number of connected vehicles with the network. The authors claimed that this technique can achieve higher accuracy, however, it was not easy to use and is better suited for more complex tasks [45].

The random forest algorithm was used for the detection of cyber-attacks in the smart cities and this algorithm was reported to achieve the accuracy of 99.34% [15]. In another study, the random forest algorithm was used to predict the charging requirements of electrical vehicles. The proposed system reportedly optimized the charging demands, however, the algorithm was complex and not easy to implement [46]. The random forest algorithm was also used in the education sector in one study. In this study, the RF algorithm was used for the classification of the development of the knowledge profile of the students. As per results, the model is suitable and it gave reasonable results [47]. One study used the RF algorithm for sustainability of the environment. S. Benedict used the RF algorithm for the estimation of air pollution parameters in the air of the smart cities. The study achieved accuracy of between 70-90%, however, other machine learning algorithms were reported to perform better than RF algorithm. Another study used the RF algorithm for the visual semantic decision support system and considering comparing RF with other algorithms, they reported that RF algorithm achieved accuracy of around 91% [48].

Although the reported studies report reasonable accuracy, they mostly fail to consider that IOT devices have limited resources. Moreover, most of them are connected via band-limited communication media and they are often designed to consume less energy. This renders these devices to have limited computation power. Apropos this, it is crucial to develop a model, which not only provides good accuracy, but is also fast and does not consume a lot of computational and memory resources. With this motivation, we aspire to design a pipeline that is fast, yet it gives high accuracy, increasing its applicability in diverse situations across IOT-based smart cities.

7. PROPOSED MODEL

7.1 Architecture

Before testing the machine learning algorithms on the data, it is pertinent to create a model for such an analysis. The pipeline for the analysis is shown in figure 3. Our pipeline consists of first using the DT algorithm for the binary classification of the data to find whether there is an attack or not. Afterwards, we will only use the data pertaining to attack and we will use RF to find the types of the attacks. Our devised model is computationally fast, as the amount of data fed to the RF algorithm will be less.

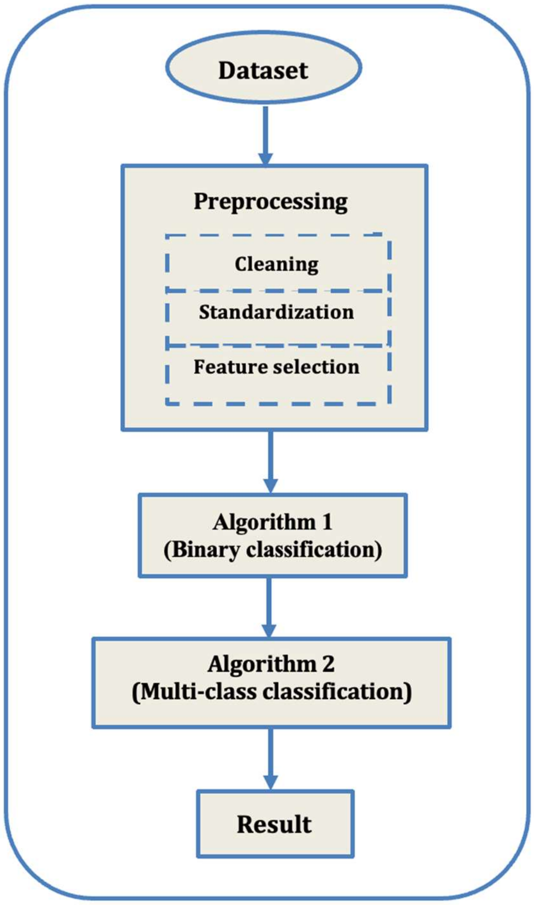

Figure 3: Proposed pipeline for the analysis of machine learning algorithms. Any dataset can be used in this model

The first step is to preprocess the data. This involves cleaning and removing unwanted values like NaNs. Moreover, the data needs to be standardized as well and this can be done by scaling. Afterwards, the features need to be selected. Not all features are always relevant, and this makes features selection an important step. The greater number of features improve accuracy of the model; however, they also increase the computational complexity of the model. Next step is to choose the appropriate classifier for the model. For our current study, we chose two classifiers, namely, decision tree classifier and random forest classifier. Decision tree is a simple machine learning algorithm; however, the random forest classifier comes under the category of advanced machine learning algorithms. Once the classifier is decided, then the next step is to train the classification model. For this usually a part of the same dataset is taken and fed to the model for the training. Both the features and the independent variable are fed to the model. Finally, the model is tested by giving it the remaining part of the dataset as well. The results of the testing of the model are assessed based on metrics described in section 6. For the current study, we analyzed the publicly available IOT dataset, TON-IOT.

Here is the pipeline approach detailed in 6 steps:

1. Preprocess: cleaning and standardization of data
2. Classification and select feature
3. Train the dataset with the DT algorithm, and the RF algorithm.
4. Test the dataset file with the DT algorithm to get an output file containing attacks only.
5. Adjust the output file to add all features
6. Test the attacks file result from precedent step, with RF algorithm.

7.2 Dataset

The dataset to be analyzed was TON-IoT (UNSW-Io20), and we choose two files: ‘IoT_Weather.csv’ and “IoT_GPS_Tracker.csv”. This data is combination of IOT and IIOT (industrial IOT), which has been gathered to check the integrity of different IOT-based applications using artificial intelligence. The dataset was collected from the School of Engineering and Information Technology (SEIT) at UNSW Canberra. The datasets were collected using parallel processing to gather several normal and cyber-attack scenarios.

Afterwards, we selected the following features, in each file, ‘date’, ‘temperature’, ‘pressure’ and ‘humidity’ for ‘IoT_Weather.csv’ and date’, ‘Latitude’, ‘Longitude’ for “IoT_GPS_Tracker.csv”. The independent variables in these files were named as ‘type’ and ‘label. The column ‘label’ only contained the binary information whether there is an attack or not, while the column ‘type’ consisted of the type of cyber-attack, and it consists of the following 7 entries, normal, backdoor, ddos, injection, normal, password, ransomeware and xss. Except ‘normal’, all others were type of cyber threats. The number of data points in all three features was 587076.
An example of traffic in Weather.csv file is given in Figure 4 below.

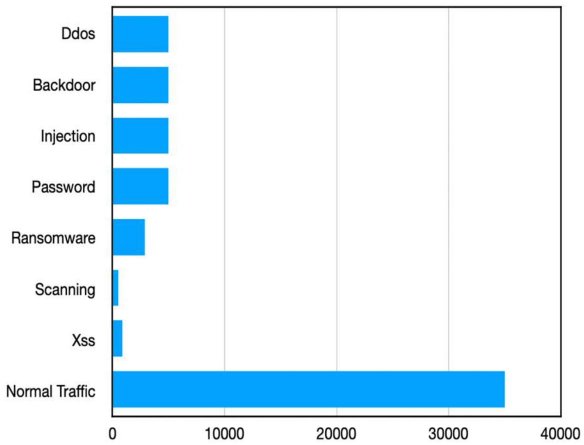

Figure 4: Example of traffic in weather.csv file dataset

7.3 Metrics for performance evaluation

To quantify the performance of any ML algorithm, several assessment metrics are used. Some of them are more common, like accuracy, precision, etc., while some like MAPE are relatively uncommon or study specific. In the next section, we will briefly describe some of the most common criterions for evaluating the performance of ML algorithms.

7.3.1 Accuracy

The classification accuracy in the context of machine learning is the ratio of sum of true positives (TP) and true negatives (TN) divided by the sum of TP, TN, false positives (FP) and false negatives (FN) [42]. The corresponding equation can be given as equation 1.

$$Accuracy = \frac{TP+}{TP+TN+FP+FN} \quad (1)$$

Accuracy is one of the most common measures of the performance of the ML algorithm. We can see those studies, which employ a single ML algorithm for classification usually report reasonable accuracy values. However, studies which merge more than one ML measures in order to improve the inference power report higher accuracy values.

7.3.2 Precision

Precision is the measure of the quality of the correct prediction made by the model. Mathematically, it can be described as the ratio of the TP by the sum of the TP and FP [42]. The equation for the precision can be given as,

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

The high value of precision means that the model has less false positives.

7.3.3 Classifier Recall

Classifier recall is the ability of the model to find all the pertinent cases within the data set. Mathematically, it is the ratio of the TP with the sum of the TP and FN [42],

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

The recall is also termed as the sensitivity and higher value of sensitivity means that model is accurately identifying maximum number of positive results.

7.3.4 F-Measure/F-Score

Precision and recall can both be combined to give a metric termed as F-measure known also as F-Score. F-measure is the weight means of precision and recall [42]. The formula for the calculation of F-measure is given as,

$$F\ measure = \frac{Precision\ X\ Recall}{Precision+Rec} \qquad (4)$$

The numerical value of F-measure is between 0 and 1, where 1 represents the best, while 0 corresponds to the worst.

7.3.5 Computation time

Computation time in machine learning refers to the amount of time it takes for a machine learning algorithm to train a model on a given dataset and/or make predictions on new data points. It is an important factor to consider when selecting and implementing machine learning algorithms, as it can significantly impact the overall efficiency and feasibility of the project.

Computation time can be influenced by several factors, like Algorithm complexity, Dataset size, Feature dimensionality, Hardware resources, Model hyperparameters [42].

8. RESULTS AND DISCUSSIONS

As the result of 70-30 division of the data, 70% of the data was used for the training, while 30% was used for the testing of the classifier. The dimensions of training data in case of TON-IoT data 'Weather.csv' as 410953 X 3, while the dimensions of testing data were 176123 X 3. We fixed the random state value to 44. For the analysis, the custom code was written in python language and an pen source, commercially available library 'sci-kit learn' was used (https://scikit-learn.org/stable/). Using the simple machine learning algorithm of Decision Tree, we achieved the classification accuracy of 99,82%. for the TON-IoT data, while considering *'label'* as an independent variable. We used DT for this binary classification as DT is fast and for such classification it gave good enough accuracy. Afterwards, we only used dataset that corresponded to actual attack, i.e., data pertaining to *'label'* value of 1. This reduced the dimensions of data and now we used Random Forest classifier for this multi-class classification. Such reduction of data significantly improved the speed of the RF algorithm, while still giving a reasonable accuracy of 99,39% for the 'Weather.csv' file and 99,82 for the 'GPS_Tracker.csv' file, and com

The accuracies are reported in table 1 for the 'weather.csv' and table 2 for the 'GPS_Tracker.csv'.

Table 1: Accuracies achieved from using our pipeline mode for the file 'Weather.csv'.

Sr. No.	Metric	Decision tree	Random forest
		With 'label' as independent variable	With 'type' as independent variable
1	Accuracy %	99.82	99.92
2	Computation time	1.34 seconds	6,71 seconds
3	Recall	99.87%	99.39%
4	Precision	99.92%	99.03%
5	F-score	99.87%	99.07

Table 2: Accuracies achieved from using our pipeline mode for the file 'GPS_Tracker.csv'.

Sr. No.	Metric	Decision tree	Random forest
		With 'label' as independent variable	With 'type' as independent variable
1	Accuracy %	98.76	98.95
6	Computation time	1.39 seconds	6,78 seconds
2	Recall	99.87%	99.39%
3	Precision	99.92%	99.03%
4	F-score	99.87%	99.07

Table 3: Comparison with other studies

Reference	Algorithm	Accuracy %	Train time	Test time
[50]	CNN-LSTM	98,02	--	--
[52]	CFBPNN	97,3	24min	--
[51]	LSTM	88	1596s	9s
Proposed approach	**DT and RF**	**>98,7**	**<1,3s**	**6,7**

Results show that both the algorithms can successfully detect cyber in the case of IOT datasets. Most of the machine learning algorithms focus on improvement of accuracy, however, in the case of IOT devices, computation complexity is also a very important factor. While designing the strategy to safeguard the often resource-restrained IOT devices, it is pertinent to take care about the computational cost of the technique employed. Our proposed pipeline approach divides the dataset and first uses the 'label' of the data set to successfully isolate the attack activity from the normal one using decision tree algorithm. Once attacks are classified, in the next step, we use random forest, which is combination of multiple trees, to identify the type of attack using 'type.' as an independent variable. Our proposed scheme reduces the amount of data to be used by the RF algorithm and hence makes it computationally fast, while offering reasonable accuracy. We advocate that our proposed model is fast and while devising schemes to ensure security and privacy in the IOT smart city, the speed of algorithms should also be given weightage along with its accuracy. The IOT network is resource-limited, hence smart, fast, and efficient algorithms should be used.

A comparison of existing works on TON-IOT dataset, that utilize intrusion detection systems for the Internet of Things (IoT) can be found in table 3, which serves as a summary for reference. The long short-term memory (LSTM) models are a class emerging models that have been widely applied on diverse class of datasets. However, they are complicated and require more data to be trained effectively. They often compromise accuracy if the input data is not in the form of a sequence. Moreover, they take lot of time to be trained on big dataset. Similarly, the feed forward and propagation-based neural networks are well suited to very big datasets. The have high computational cost and are prone to the problem of having unstable parameters, which decreases their efficiency and effectiveness.

Some of the limitations of our study might include the fact that decision tree algorithms are very data sensitive. A small change in dataset might result in totally different results than expected. Furthermore, our method is preferred only in the case where the number of features in the dataset is high enough. For the random forest algorithm, the dataset needs to have a certain degree of randomness to work efficiently. Moreover, we also did not carry out much feature engineering in the present study. We believe that feature engineering will further improve the results, however, such exercise might also include the bias towards features.

9. CONCLUSIONS:

Increased urbanization will require cities to be smart not to cope but also to offer a better standard of living. The IOT-based smart cities will not only enhance quality of life, but they will also use the limited resources intelligently and efficiently. To do so, continuous monitoring of resources is vital and for that a lot of devices need to be connected to the internet in the smart city. Such communication of data poses a risk of cyber-attacks. In the presented paper, we have proposed a pipeline that can be used to safeguard the IOT data. Our proposed model is computationally fast, yet it gives reasonable accuracy. Moreover, several simple and advanced machine learning algorithms were described as well, that have the capacity to mitigate cyber-attacks and threats. Lot of machine learning literature targets improving the accuracy of the models and literature is full of very accurate models, however, what needs to be considered is the fact that most of the IOT devices have very limited resources their disposal. Whether the IOT device will be able to run the model is important as well. This study focused on the practicality and applicability of the machine learning model. We highlighted that in addition to accuracy, the computational complexity and speed of the model is also important, perhaps, at times even more important. Our study also presents an emerging and important avenue of the application of machine learning. Smart cities are the future of urban housing. Our study strengthens the fact that AI is crucial in improving humans lives in our resource-depleted environment. Particularly important is the role of artificial intelligence in managing huge amounts of IOT data, but also ensuring its effective protection and storage.

The discussion and results presented in this paper will open the way for further research for the development of more robust and accurate ML algorithms for the privacy and security of data in smart cities. We also aspire that findings of this paper will contribute for the cyber security research of IOT-based smart cities. Further work must be carried out to improve the performance of already existing machine learning and deep learning algorithms.

REFERENCES:

[1] H. M. K. K. M. B. Herath and M. Mittal, "Adoption of artificial intelligence in smart cities: A comprehensive review," *International Journal of Information Management Data Insights,* vol. 2, no. 1, p. 100076, 2022/04/01/

2022, doi: https://doi.org/10.1016/j.jjimei.2022.100076.

[2] U. J. D. o. E. Nations and P. D. Social Affairs, World Urbanization Prospects, "The World's cities in 2018," pp. 1-34, 2018.

[3] R. Dowling, P. McGuirk, C. J. U. P. Gillon, and Research, "Strategic or piecemeal? Smart city initiatives in Sydney and Melbourne," vol. 37, no. 4, pp. 429-441, 2019.

[4] A.-M. D. Adunadepo and O. Sunday, "Artificial intelligence for sustainable development of intelligent buildings," in *Proceedings of the 9th CIDB Postgraduate Conference, Cape Town, South Africa*, 2016, pp. 1-4.

[5] N. Kapoor, N. Ahmad, S. K. Nayak, S. P. Singh, P. V. Ilavarasan, and P. J. I. J. o. I. M. D. I. Ramamoorthy, "Identifying infrastructural gap areas for smart and sustainable tribal village development: A data science approach from India," vol. 1, no. 2, p. 100041, 2021.

[6] F. J. F. i. S. C. Cugurullo, "Urban artificial intelligence: From automation to autonomy in the smart city," vol. 2, p. 38, 2020.

[7] K. Haseeb, N. Islam, Y. Javed, and U. Tariq, "A Lightweight Secure and Energy-Efficient Fog-Based Routing Protocol for Constraint Sensors Network," vol. 14, no. 1, p. 89, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/1/89.

[8] H. Yar, T. Hussain, Z. A. Khan, D. Koundal, M. Y. Lee, and S. W. Baik, "Vision Sensor-Based Real-Time Fire Detection in Resource-Constrained IoT Environments," *Comput Intell Neurosci,* vol. 2021, p. 5195508, 2021/12/21 2021, doi: 10.1155/2021/5195508.

[9] T. Saba, T. Sadad, A. Rehman, Z. Mehmood, and Q. Javaid, "Intrusion Detection System Through Advance Machine Learning for the Internet of Things Networks," *IT Professional,* vol. 23, no. 2, pp. 58-64, 2021, doi: 10.1109/MITP.2020.2992710.

[10] S. Chakrabarty and D. W. Engels, "Secure smart cities framework using IoT and AI," in *2020 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, 2020: IEEE, pp. 1-6.

[11] M. A. Rahman *et al.*, "Scalable machine learning-based intrusion detection system for IoT-enabled smart cities," vol. 61, p. 102324, 2020.

[12] S. Selvaganapathy, M. Nivaashini, and H. J. I. S. J. A. G. P. Natarajan, "Deep belief network based detection and categorization of malicious URLs," vol. 27, no. 3, pp. 145-161, 2018.

[13] H. Zhang, Y. Li, Z. Lv, A. K. Sangaiah, and T. J. I. C. J. o. A. S. Huang, "A real-time and ubiquitous network attack detection based on deep belief network and support vector machine," vol. 7, no. 3, pp. 790-799, 2020.

[14] J. A. Nasir, O. S. Khan, and I. J. I. J. o. I. M. D. I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," vol. 1, no. 1, p. 100007, 2021.

[15] I. Alrashdi, A. Alqazzaz, E. Aloufi, R. Alharthi, M. Zohdy, and H. Ming, "AD-IoT: Anomaly Detection of IoT Cyberattacks in Smart City Using Machine Learning," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 7-9 Jan. 2019 2019, pp. 0305-0310, doi: 10.1109/CCWC.2019.8666450.

[16] S. S. Band *et al.*, "When smart cities get smarter via machine learning: An in-depth literature review," 2022.

[17] M. Elhoseny, K. Haseeb, A. A. Shah, I. Ahmad, Z. Jan, and M. I. Alghamdi, "IoT Solution for AI-Enabled PRIVACY-PREServing with Big Data Transferring: An Application for Healthcare Using Blockchain," vol. 14, no. 17, p. 5364, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/17/5364.

[18] R. Abbasi, L. Xu, F. Amin, and B. Luo, "Efficient Lossless Compression Based Reversible Data Hiding Using Multilayered n-Bit Localization," *Security and Communication Networks,* vol. 2019, p. 8981240, 2019/01/15 2019, doi: 10.1155/2019/8981240.

[19] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems," *IEEE Access,* vol. 8, pp. 165130-165150, 2020, doi: 10.1109/ACCESS.2020.3022862.

[20] N. Islam, M. Altamimi, K. Haseeb, and M. Siraj, "Secure and Sustainable Predictive Framework for IoT-Based Multimedia Services Using Machine Learning," *Sustainability,* vol. 13, p. 13128, 11/26 2021, doi: 10.3390/su132313128.

[21] K. Haseeb, I. Ahmad, I. Awan, J. Lloret, and I. Bosch, "A Machine Learning SDN-Enabled Big Data Model for IoMT Systems," *Electronics,* vol. 10, 09/11 2021, doi: 10.3390/electronics10182228.

[22] B. Z. H. Zhao, M. Ikram, H. J. Asghar, M. A. Kaafar, A. Chaabane, and K. Thilakarathna, "A Decade of Mal-Activity Reporting: A Retrospective Analysis of Internet Malicious Activity Blacklists," presented at the Proceedings of the 2019 ACM Asia Conference

on Computer and Communications Security, Auckland, New Zealand, 2019. [Online]. Available: https://doi.org/10.1145/3321705.3329834.

[23] K. Haseeb, N. Islam, A. S. Almogren, and I. J. I. A. Ud Din, "Intrusion Prevention Framework for Secure Routing in WSN-Based Mobile Internet of Things," vol. 7, pp. 185496-185505, 2019.

[24] Y. Zheng, Z. Xu, and A. Xiao, "Deep learning in economics: a systematic and critical review," *Artificial Intelligence Review,* 2023/02/04 2023, doi: 10.1007/s10462-022-10272-8.

[25] Y. Freund and L. Mason, "The Alternating Decision Tree Learning Algorithm," presented at the Proceedings of the Sixteenth International Conference on Machine Learning, 1999.

[26] R. AlZoman and M. Alenazi, "A Comparative Study of Traffic Classification Techniques for Smart City Networks," *Sensors,* vol. 21, p. 4677, 07/08 2021, doi: 10.3390/s21144677.

[27] D. Ngabo, W. Dong, E. Ibeke, C. Iwendi, and E. Masabo, "Tackling pandemics in smart cities using machine learning architecture," (in eng), *Mathematical biosciences and engineering : MBE,* vol. 18, no. 6, pp. 8444-8461, Sep 27 2021, doi: 10.3934/mbe.2021418.

[28] M. Shen, X. Tang, L. Zhu, X. Du, and M. Guizani, "Privacy-Preserving Support Vector Machine Training Over Blockchain-Based Encrypted IoT Data in Smart Cities," *IEEE Internet of Things Journal,* vol. 6, no. 5, pp. 7702-7712, 2019, doi: 10.1109/JIOT.2019.2901840.

[29] M. Rashid, J. Kamruzzaman, M. Hassan, T. Imam, and S. Gordon, "Cyberattacks Detection in IoT-Based Smart City Applications Using Machine Learning Techniques," *International Journal of Environmental Research and Public Health,* vol. 17, p. 9347, 12/14 2020, doi: 10.3390/ijerph17249347.

[30] D. Protic, L. Gaur, M. Stankovic, and M. A. Rahman, "Cybersecurity in Smart Cities: Detection of Opposing Decisions on Anomalies in the Computer Network Behavior," vol. 11, no. 22, p. 3718, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/22/3718.

[31] O. Said, A. J. S. C. Tolba, and Society, "Accurate performance prediction of IoT communication systems for smart cities: An efficient deep learning based solution," vol. 69, p. 102830, 2021.

[32] E. M. Onyema, S. Dalal, C. A. T. Romero, B. Seth, P. Young, and M. A. Wajid, "Design of Intrusion Detection System based on Cyborg intelligence for security of Cloud Network Traffic of Smart Cities," *Journal of Cloud Computing,* vol. 11, no. 1, p. 26, 2022/08/13 2022, doi: 10.1186/s13677-022-00305-6.

[33] N. Al-Taleb and N. A. Saqib, "Towards a Hybrid Machine Learning Model for Intelligent Cyber Threat Identification in Smart City Environments," vol. 12, no. 4, p. 1863, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/4/1863.

[34] S. Faizollahzadeh Ardabili, B. Najafi, M. Alizamir, A. Mosavi, S. Shamshirband, and T. Rabczuk, "Using SVM-RSM and ELM-RSM Approaches for Optimizing the Production Process of Methyl and Ethyl Esters," vol. 11, no. 11, p. 2889, 2018. [Online]. Available: https://www.mdpi.com/1996-1073/11/11/2889.

[35] J. Dou *et al.*, "Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan," *Landslides,* vol. 17, no. 3, pp. 641-658, 2020/03/01 2020, doi: 10.1007/s10346-019-01286-5.

[36] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapè, "Multi-Classification Approaches for Classifying Mobile App Traffic," *Journal of Network and Computer Applications,* vol. 103, 11/14 2017, doi: 10.1016/j.jnca.2017.11.007.

[37] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Mobile Encrypted Traffic Classification Using Deep Learning: Experimental Evaluation, Lessons Learned, and Challenges," *IEEE Transactions on Network and Service Management,* vol. 16, no. 2, pp. 445-458, 2019, doi: 10.1109/TNSM.2019.2899085.

[38] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach," *Procedia Computer Science,* vol. 132, pp. 679-688, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.procs.2018.05.069.

[39] W. Hao, W. Yizhou, L. Yaqin, and S. Zhili, "The Role of Activation Function in CNN," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, 18-20 Dec. 2020 2020, pp. 429-432, doi: 10.1109/ITCA52113.2020.00096.

[40] M. Mohammadpourfard, A. Khalili, I. Genc, and C. Konstantinou, "Cyber-Resilient Smart Cities: Detection of Malicious Attacks in Smart Grids," *Sustainable Cities and Society,* vol. 75,

p. 103116, 2021/12/01/ 2021, doi: https://doi.org/10.1016/j.scs.2021.103116.

[41] A. Gensler, J. Henze, B. Sick, and N. Raabe, "Deep Learning for solar power forecasting — An approach using AutoEncoder and LSTM Neural Networks," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 9-12 Oct. 2016 2016, pp. 002858-002865, doi: 10.1109/SMC.2016.7844673.

[42] C. Sweeney, E. Ennis, M. Mulvenna, R. Bond, and S. O'Neill, "How Machine Learning Classification Accuracy Changes in a Happiness Dataset with Different Demographic Groups," vol. 11, no. 5, p. 83, 2022. [Online]. Available: https://www.mdpi.com/2073-431X/11/5/83.

[43] D. Ngabo, W. Dong, E. Ibeke, C. Iwendi, and E. Masabo, "Tackling pandemics in smart cities using machine learning architecture," *Mathematical Biosciences and Engineering,* vol. 18, no. 6, pp. 8444-8461, 2021, doi: 10.3934/mbe.2021418.

[44] R. Bardhan, R. Debnath, J. Gama, and U. Vijay, "REST framework: A modelling approach towards cooling energy stress mitigation plans for future cities in warming Global South," *Sustainable Cities and Society,* vol. 61, p. 102315, 2020/10/01/ 2020, doi: https://doi.org/10.1016/j.scs.2020.102315.

[45] M. Aloqaily, S. Otoum, I. A. Ridhawi, and Y. Jararweh, "An intrusion detection system for connected vehicles in smart cities," *Ad Hoc Networks,* vol. 90, p. 101842, 2019/07/01/ 2019, doi: https://doi.org/10.1016/j.adhoc.2019.02.001.

[46] T. Xu, H. Sun, G. Han, C. Ma, and L. Jiang, "A Deployment Model of Charging Pile Based on Random Forest for Shared Electric Vehicle in Smart Cities," in *2018 14th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, 6-8 Dec. 2018 2018, pp. 49-54, doi: 10.1109/MSN.2018.00015.

[47] E. Gomede, F. H. Gaffo, G. U. Briganó, R. M. De Barros, and L. D. S. Mendes, "Application of Computational Intelligence to Improve Education in Smart Cities," vol. 18, no. 1, p. 267, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/1/267.

[48] N. Sideris, G. Bardis, A. Voulodimos, G. Miaoulis, and D. Ghazanfarpour, "Using Random Forests on Real-World City Data for Urban Planning in a Visual Semantic Decision Support System," (in eng), *Sensors (Basel),* vol. 19, no. 10, p. 2266, 2019, doi: 10.3390/s19102266.

[49] Tanzila Saba, " Securing the IoT System of Smart City against Cyber Threats Using Deep Learning ".

[50] Tahani Gazdar , " FDeep: A Fog-based Intrusion Detection System for Smart Home using Deep Learning ".

[51] Alsaedi, A.; Moustafa, N.; Tari, Z.; Mahmood, A.; Anwar, A. TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. IEEE Access 2020, 8, 165130–165150

[52] P.L.S. JAYALAXMI A, GULSHAN KUMAR A,B,∗, RAHUL Saha a,b, Mauro Conti b, Tai-hoon Kim c, Reji Thomas d « DeBot: A deep learning-based model for bot detection in industrial internet-of-things