Ransomware Auto-Detection In IoT Devices Using Machine Learning

Anshuman Dash
["REVA Academy for Corporate Excellence
REVA University
Bengaluru, India
anshumand.ba03@reva.edu.in"]

Satyajit Pal
["REVA Academy for Corporate Excellence
REVA University
Bengaluru, India"]
satyajitp.ba03@reva.edu.in

Girish Y
["REVA Academy for Corporate Excellence
REVA University
Bengaluru, India"]
Girish.cs01@reva.edu.in

Chinmay Hegde
["REVA Academy for Corporate Excellence
REVA University
Bengaluru, India"]
chegde@astrikosconsulting.com

Abstract— Internet of Things (known as IoT) "is an ecosystem of connected physical objects that are accessible through the internet". The IoT are tiny devices containing sensors that are connected using wired or wireless networks. IoT technology has grown quickly during the last decade. However, security of IoT devices have been are the weakest areas. "There are over six billion estimated devices currently connected to the Internet and an estimate of over 25 billion will be connected by 2020". The technologies of IoT and various applications in real life has spread from health care and food producing industries to the building of smart cities and urban management.

While the application of IoT has improved efficiency in the recent days, security issues have become a concern for the industry using these services. As IoT devices are always connected to the internet, they have become easy target for cyber criminals because of their generality and the possibility to use the compromised device to attack the architecture underlying these technologies. Devices that can store reasonable amount of data are generally targeted by ransomware. Hence it has become an ongoing topic of discussion to find ways and means to ensure security of IoT devices against threats like malware. Ransomware is a type of malware that attacks and encrypts the compromised devices data and storage using relatively strong encryption algorithm, forcing the victim to pay to the attacker to get the password or decryption keys. This may result in temporary or may be permanent loss of sensitive information and may cause financial losses.

In this paper, we are presenting an approach based on machine learning to be able to detect ransomware attacks on IoT devices. Our approach includes machine learning models like "K-Nearest Neighbors, Neural Networks, Support Vector Machine and Random Forest", and we will be using measurements like accuracy percentage, recall percentage and precision percentage.

["Keywords— *Ransomware detection, Internet of Things (IoT) security, Machine learning, Malware detection, DDoS"]*

I. INTRODUCTION

The concept of Internet of things(IoT) was introduced by the growth of internet along with the deployment of ubiquitous computing and mobiles in smart objects which brings new opportunities for the creation of innovative solutions to various aspects of life [1]. The concept of Internet of things(IoT)creates a network of objects that can communicate, interact and cooperate to reach a common goal [2]. IoT devices can enhance our daily lives, as each device stops acting as a single device and become part of an entire full connected system. This provides us with the resulting data to be analyzed for better decision making, tracking our businesses and monitoring our properties while we are far away from them [3].

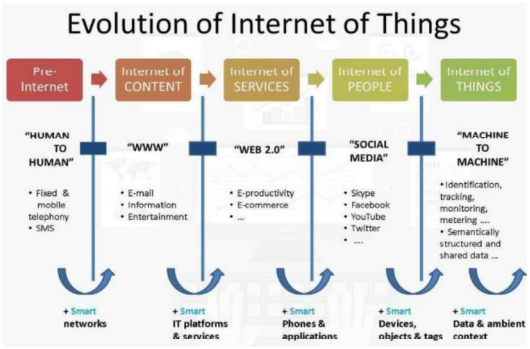


Fig. 1: Evolution of IoT
(https://twitter.com/fisher85m/status/926360908900773889)

II. IOT APPLICATIONS

As the model of IoT is growing, every part of our life is getting impacted by it. This is leading to an easier life through the wide range of applications of IoT such as smart city and its utilities.

A. *Basic Utility Systems in Smart City*

- *Smart Water Management System:* In this innovative day and age, the Internet of Things (IoT) is offering new solutions for improving water management to maximize efficient use. Comprehensive strategies utilizing the IoT can reduce water costs up to 20 percent. One of the reasons is how cheap IoT sensors are becoming with new battery-powered networking solutions (like LORA).
- *Smart Energy Management Systems:* Energy is a an extremely important aspect for our household, industries, agriculture and so on. It is hence important to Manage the energy efficiently and conserve it intelligently for various appliances. The energy usage

is directly affected with Coal, oil and so towards power generation.

- *Smart Gas Management Systems:* Urban natural gas provisioning has made significant progress in recent years. In the smart gas field, IoT enables stable, real-time traffic data collection from gas meters, device status monitoring, command delivery, and additional remote operations.
- *Smart Sewage Management Systems:* IOT in wastewater facility installs smart sensors at various points in their water management system. These sensors collect data on water quality, temperature variations, pressure changes, water, and chemical leaks, and they send those data back to a web application that synthesizes the information into actionable insights.
- *Smart Lighting Management Systems:* Improving safety within the city is one of the major benefits of smart lighting solution. The way smart lights work is that they derive power from the power grid and the street light poles have small cells that are a key enabler for smart solutions.

Fig. 2: An illustration of IoT based Smart City with Smart Utility Systems

B. *Safety & Security (Surveillance)*

We envisage our cities will have everything automated through technologies like "artificial intelligence (AI), cloud, Big Data and analytics". But the importance of 24/7, round the clock surveillance which is the key component shouldn't be forgotten. This is required to ensure the safety and security of citizens.

- *Environmental sensors:* Eenvironmental conditions like "temperature, air quality, humidity, noise levels and water quality" can be measured using sensors. This helps the authorities to allocate resources for any mitigations required.
- *Support to law enforcement:* Law enforcement is more effective when this is done through Video surveillance. This can help the agencies to identify and gather evidence during an investigations. Therefore it is import to install high quality CCTV cameras will ensure recorded video is of high quality.

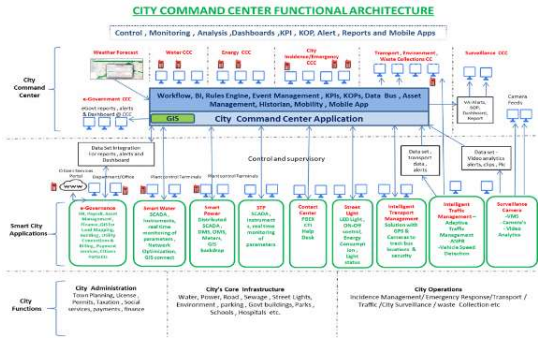


Fig. 3: An illustration of IoT based Smart City Functional Architecture

(https://docs.google.com/presentation/d/1ck3o4M21qEvvXTLgXL56mdCwi6pvYlmITXNyv7JV0TU/edit#slide=id.g3e3c360a4a_0_27)

C. *Mobility*

- *Smart Traffic Management Systems:* Traffic management has become an infrastructure hurdle for the developing countries today due to the increase in number of vehicles on the road has been resulting in traffic congestions. Various countries are trying to manage these traffic bottlenecks by using and analyzing the data fetched through CCTV camera feeds and transmitting all such data related vehicles to the city traffic management centres. This may help create improvements in the coming days.

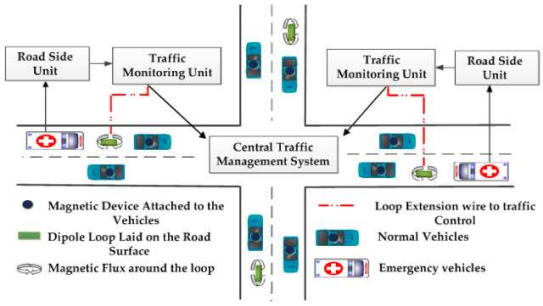


Fig. 4: An illustration of IoT based Smart Traffic Managements System

(https://www.mdpi.com/1424-8220/16/2/157)

- *Smart Transport Management Systems:* "Smart transportation system is developed over smart infrastructure that includes multi-modal connected conveyance. This also includes automated traffic signals, tolls and fare collection systems." Smart services can offer various different benefits, starting from smart parking and vehicle locating systems, and can even help in route diversion alerts.

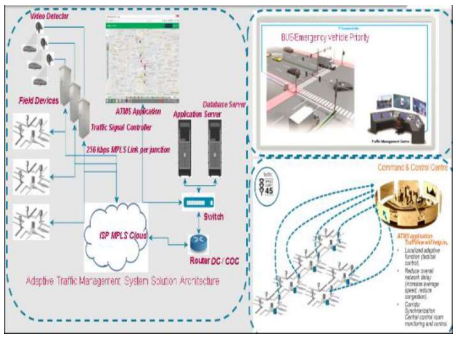


Fig. 5: An illustration of IoT based Smart Transport Management System

("https://docs.google.com/presentation/d/1ck3o4M21qEvvXTLgXL56mdCwi6pvYlmITXNyv7JV0TU/edit#slide=id.g3e3c360a4a_0_27")

D. Governance

Terms like Smart governance and good governance are like the two sides of a coin. Usage of Digital technologies and Internet is helping create a progressive government and public partnerships, it's helping strengthen government institutions and is helping integrate various sections of society. To be able to effectively manage such segments of our society, our cities need to have smart administration as well as governance. Various options like "web portals, online forums, mobile apps, and unified services" provided through these have helped people to directly share their suggestions, questions or grievances to various government authorities.

III. IoT Technologies And Protocols

Several Communication Protocols and Technology used in the internet of Things. Some of the major IoT technology and protocol (IoT Communication Protocols) are discussed here. These IoT communication protocols cater to and meet the specific functional requirement of an IoT system.

A. 6LOWPAN

6LoWPAN helps connecting more and more end point devices to the cloud. The ability to consume Low-power, and the capability to have nodes running on IP address along with the support of large mesh networks make this technology a great option for Internet of Things (IoT) applications. ["As the full name implies – IPv6 over Low-Power Wireless Personal Area Networks – 6LoWPAN is a networking technology or adaptation layer that allows IPv6 packets to be carried efficiently within small link layer frames, such as those defined by IEEE 802.15.4 [7]".]

- *Network Architecture:* The uplink to the Internet is primarily done by the Access Point (AP) which acts as an IPv6 router. Other devices such as PCs, servers, etc. then get connected to the AP. The 6LoWPAN network is connected to the IPv6 network using an edge router.
- *Security:* Security requirement is a necessity for IoT systems and have always been a challenge. Due to the nature of IoT which generally has many nodes, there are also availability more number of entry points for an outside attacker. The other important aspect is that data flowing in a typical IoT has a much higher damage potential since the data flowing in the system can be used to open the door of your house or turn on/off functionality of devices remotely.

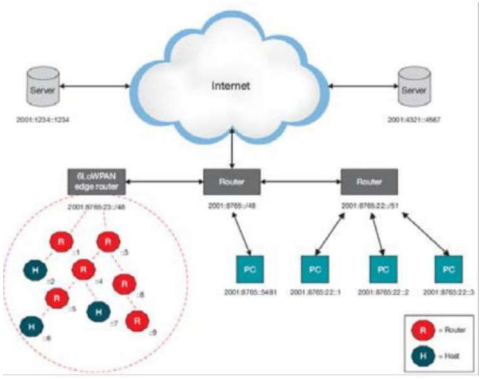


Fig. 6: 6LoWPAN Network Architecture

("https://www.researchgate.net/figure/An-example-of-an-IPv6-network-with-a-6LoWPAN-mesh-network-26_fig2_318075856")

B. "DASH7":

"DASH7 Alliance protocol is an Actuator network protocol and it is an open source wireless sensor network. DASH7 operates in 433 MHz, 868 MHz and 915 MHz unlicensed ISM bands." DASH7 can generally provide a range up to 2km. Security on DASH7 is based on AES 128-bit shared key encryption. Data rate offered by DASH7 is "28 kbps" and it uses "wakeup signal to achieve low power, low latency and elegant architecture".

"DASH7 uses BLAST networking technology. BLAST means Bursty Light data Asynchronous and Transitive." "In bursty, data transfer is abrupt and it does not contain contents like audio and video." "In light, multiple consecutive packet transmission is generally avoided and the packet size is limited to 256 bytes." DASH7 doesn't need periodic hand shaking as it communicates through command responses.

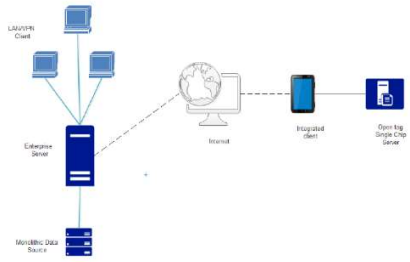


Fig. 7: DASH7 Network Architecture

(http://iopscience.iop.org/article/10.1088/1757-899X/396/1/012027/pdf)

C. "LoRa & LoRaWAN":

- *"LoRa":* Semtech Corporation has made a "radio modulation technology". LoRa [3] can support connectivity of about 15 to 20 km using "chirp spread spectrum technique" where the whole bandwidth is used to transmit signals. LoRa [8] uses "868 MHz to 900 MHz ISM bands for its operation and data rate is 0.3 kbps". As LoRa has a longer battery life, cost of replacing devices is much lower and its deployment is rather simple. LoRa uses "symmetric key cryptography" to ensure security to its devices.

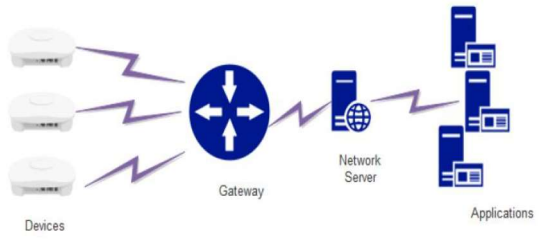


Fig. 8: LoRa Network Architecture

(http://iopscience.iop.org/article/10.1088/1757-899X/396/1/012027/pdf)

- *"LoRaWAN":* The communication protocol and system architecture for the network is defined by LoRaWAN The physical layer of LoRa helps with the long range communication. Battery life of any particular node depends on what protocol and network architecture is used. It will also depend on the capacity of network, the service quality, security etc that is provided by the network.
- *"LoRaWAN" Network Architecture:* most of the existing deployed networks utilize a mesh network architecture. "In a mesh network, the individual end-nodes forward the information of other nodes to increase the communication range and cell size of the network." This helps increase the range, and also adds complexity, reduce network capacity and reduces battery lifetime. Long range architectures are the best when it comes preserving battery life while providing long range connectivity.
- *Security:* "LoRaWAN™" uses two different layers of security. One is for Network and the other for the Application. Node authenticity is verified by network security and the end user's data within an application is secured using the application layer security. "AES encryption is used with the key exchange utilizing an IEEE EUI64 identifier".

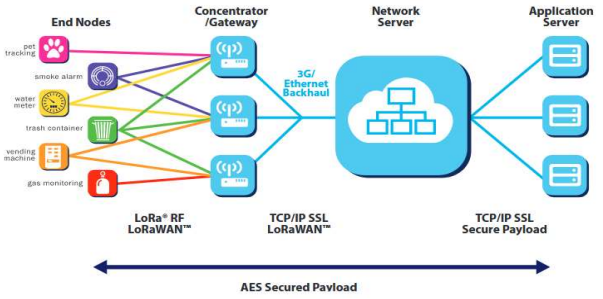


Fig. 9: "LoRaWAN Network Architecture"

(https://www.profissionaisti.com.br/2017/11/iot-protocolo-lorawan-e-principais-placas-de-desenvolvimento-lora/)

D. SigFox

"Sigfox is a LPWAN technology which provides low power low data and low-cost communication devices. It operates in global network and is used for IoT communications." Even though there are several wireless technologies existing, "Sigfox" is widely accepted as it provides cheaper connectivity options and it helps extend battery life. "Sigfox" can provide connectivity for devices with higher bandwidth while enabling newer IoT applications.

- *SigFox Network Architecture:* "Communication over Sigfox can be performed when it detects an event or measure something, it will power on the communication module and send the message. The message is then picked up by the network and the data is received on your server". Sigfox is enhances energy efficiency as it consumes very low power while transmiting data and doesn't need any maintenance.
- *Security:* Systematic process can be used to address security challenges in "Sigfox". "Sigfox is one of the most secured LPWAN technology and it is unique in design". Sigfox devices principally operate in offline mode with a defined behavior. Sigfox device would start broadcasting radio messages when it must connect to internet for the purpose of transmission of data. This message broadcast is then intercepted by the base stations and Sigfox core network received the transmitted messages, which in turn gets delivered to the matching IoT applications. The network architecture usually offers an air gap and it cannot be accessed through internet maliciously.

Fig. 10: SigFox Network Architecture

(http://www.rfwireless-world.com/Tutorials/Sigfox-network-architecture.html)

E. Zigbee

["Zigbee system structure consists of three different types of devices such as Zigbee coordinator, Router and End device. Every Zigbee network must consist of at least one coordinator which acts as a root and bridge of the network. The coordinator is responsible for handling and storing the information while performing receiving and transmitting data operations".]

- *Zigbee Network Architecture: The* routers of Zigbee act as midway devices that permit data to pass to and from other devices. "The number of routers, coordinators and end devices depends on the type of network such as star, tree and mesh networks".
- *Security:* The "ZigBee" standard uses compound security procedures when it comes to security. And the services are generally applied to the "Network and the Application Support Sublayer (APS), a sub layer of the Application Layer". The ZigBee protocol is based on an "open trust" model. It means there is trust relationship between various protocol stack layers. This ensures there is "cryptographic protection" between the devices. Each of the layer becomes accountable for the security of required frames.

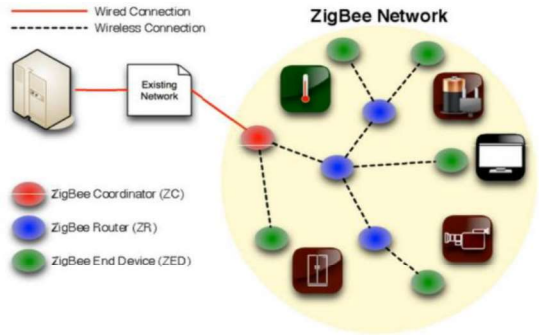


Fig. 11: Zigbee Network Architecture

(https://www.elprocus.com/what-is-zigbee-technology-architecture-and-its-applications/)

IV. RANSOMWARE

A ransomware attack is traditionally different from normal malware threats and when it comes to IoT devices, it can have devastating affect on an entire landscape of security services like "confidentiality, integrity, and availability" (the CIA), which in turn may result in financial losses and important information breach [5]. Ransomware generally takes control of data or the whole system. It may ask for a hefty sum as a ransom and may agree to let the user get its data back after payment of demanded sum. In case the user fails to pay, ransomware may delete the data from the devices or may make it permanently corrupted [6].

["Initially, ransomware was named "AIDS", as reported in 1989, when Joseph Popp distributed 20,000 infected floppy disk drives to the participants of World Health Organizations' AIDS conference [8]".] AIDS monitored the systems and calculated the system reboot count. It generally does encrypt data files and/or make the folders and directories hidden in C drive of the infected computer. AIDS was virtually undetectable as it stayed silently in the systems and could get activated after a system reboots for "90 times".

A. *Crypto Ransomware*

- "Crypto-ransomware is a type of harmful program that encrypts files stored on a computer or mobile device to extort money. Encryption 'scrambles' the contents of a file, so that it is unreadable." To restore the encrypted files, a decryption key is needed to "unscramble" the file. Crypto-ransomware takes the user data hostage and demands a ransom in exchange of the decryption key needed to restore the files.
- Crypto-ransomware prominently displays lurid messages to call attention, and openly uses shock and fear to pressure the user into paying the ransom. Some of the so-called crypto-ransomware do not even encrypt data at all, and just use the threat methodology to extort money.

B. *Locker Ransomware*

- "Locker ransomware is a virus that infects PCs and locks the user's files, preventing access to data and files located on the PC until a ransom or fines are paid." Locker is seen demanding payment of $150 via "Perfect Money or is a QIWI Visa Virtual Card" number to unlock files. The effected systems by this variant includes Windows XP, Windows Vista, Windows 7, and Windows 8. Locker ransomware is a copy of another very powerful ransomware that has infected over "250,000" computer systems and is named "Crypto Locker". Although it may seem that the Locker ransomware is simple, it can have far devastating results to user's computer.

C. *Hybrid Ransomware*

- "Hybrid ransomware attacks that enable encryption and locking mechanisms are more dangerous because the device data and functionality could be compromised." A hybrid ransomware attack could be more dangerous to IoT devices and systems because it can target front-end and back-end of the devices.

V. MACHINE LEARNING ALGORITHMS

We will first discuss about machine learning as a field and then we would describe the methods used in this study. In this paper, we are presenting an approach based on machine learning to be able to detect ransomware attacks on IoT devices. Our approach includes machine learning models like K-Nearest Neighbors, Neural Networks, Support Vector Machine and Random Forest, and we will be using measurements like accuracy rate, recall rate and precision rate. The rapid growth and expansion of data mining techniques and methodologies have resulted in "Machine Learning" creating a separate field of Computer Science. The methodology of machine learning task is to use data to train the model, based on one or more algorithms, to perform a certain type of tasks, like classification, regression, clustering etc. Training is primarily done by using a set of input datasets, and the model that is built is then used to make predictions.

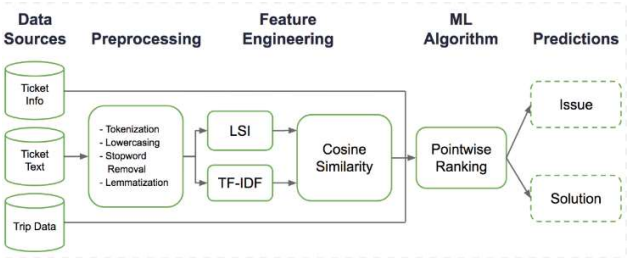


Fig. 12: General Workflow Of The Machine Learning Process

(https://eng.uber.com/cota/)

A. *Supervised and Unsupervised Learning*

- Machine learning has two types of approaches named as "supervised and unsupervised" learning. In case of "Supervised Learning", learning is defined based on labelled data. In our case, we are using a data set that has samples mapped to the correct outcome. The model very clearly knows which field is representing the correct results. in case of "Unsupervised Learning", there is no initial labelling of data. The goal of unsupervised learning is to find patterns in the unsorted data set and not to predicting some value.

B. *Classification Methods*

- Using machine learning for Ransomware detection is seen as a classification problem. While unknown Ransomware types can be cauterized into several clusters, which would be based on certain types of properties identified by the algorithm. But if we are able to train a model containing dataset of malicious and benign files, this can be converted to a problem of classification. While identifying known Ransomware

families, we can narrow down to classification only – having a limited set of classes and the ransomware can certainly belong to one of such families and it may become easier to identify proper class. The results in this case would be more accurate than with cauterization algorithms.

- *K-Nearest Neighbors*: K-Nearest Neighbors (KNN) can be called as one of the simplest and accurate machine learning algorithms. KNN is "Non-parametric" algorithm, it means it does not make any assumptions about the data structure. Both classification and regression problems can be solved using KNN. The prediction in KNN is based on the k training instances that are closest to the input instance. When we speak about KNN classification, the output would be of the same class as the input class and predicted by the "Majority vote of the k closest neighbors".

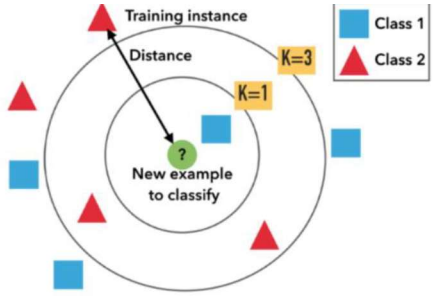


Fig. 13: KNN Example

(https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7)

$$Hamming\ Distance:\ d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

$$Manhattan\ Distance:\ d_1(p,q) = ||p-q||_1 = \sum_{i=1}^{n} |p_i - q_i|$$

$$Minkowski\ Distance = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

- Euclidean Distance is mostly used for continuous variables and is defined by the formulae below:

$$EuclidianDistance = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}\ ;\ p\ and\ q\ are\ the\ points\ in\ n-space$$

- Where ever the features of the data sets are of the same type, "Euclidian distance" is a good model for the problems. The k value plays a crucial role in the prediction accuracy of this algorithm. Smaller K values may give less accurate results, specifically in the datasets that include much noise. Selection of k is ideally done using the formula below:

$$k = \sqrt{n}$$

- *"Support Vector Machines"*: "Support Vector Machines (SVM) is another machine learning algorithm that is generally used for classification problems". The primary idea of SVM is to find such a "Hyperplane", that separates the classes in the best way. "Support vectors" are a term used to refer to the points that are closer to the "hyperplane", that would change it's position if removed. The term margin is used to refer to the distance between the support vectors and the hyperplane.

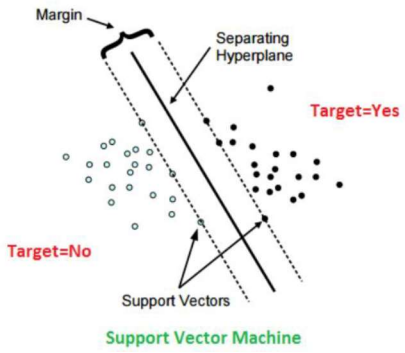


Fig. 14: SVM Example

(http://www.1proga.info/support-vector-machine-algorithm-389128593a10a109ec0dd4/)

- In the figure given above,the data set seen contains two classes. It means there is a problem in a "two dimensional" space, and then the "hyperplane" gets constructed as a line.
- The algorithm is described as below:
  - "*X* and *Y* are defined as the input and output sets respectively. ($x1$,$y1$),… upto (xm,ym) is the training set."
  - "Based on x, we want to predict y. We can define this as y=f(x, a), where a is the parameter of the classification function."
  - "F(x, a) can be learnt by minimizing the training error of the function that needs to learn on the training data. L here is defined as the loss function, and *Remp* is referred as empirical risk."

$$R_{emp}(a) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i, a), y_i) = Training\ Error$$

- We also need to now reduce the overall risk. Here, "P(x,y)" is the joint distribution function of "x and y".

$$R(a) = \int l(f(x,a),y)dP(x,y) = Test\ Error$$

- We then need to reduce the training Error along with the complexity term. Hence, we select the set of "hyperplanes", so "f(x) = (w·x)+b"

$$\frac{1}{m}\sum_{i=1}^{m} l(w \cdot x_i + b, y_i) + ||w||^2 \text{ subject to } min_i|w \cdot x_i| = 1$$

- SVMs have the ability to produce good accuracy results, specifically on "clean" datasets. It is also good

while with working with high-dimensional datasets, where dimensions are more compared to the size of the samples. Though, for large datasets with lots of noise or overlapping classes, SVMs can be proved to be more effective.

- *Random Forest:* "Random Forest" is a very popular algorithm in machine learning. It hardly requires any data preparation and modelling while resulting in accurate results. In general, "Random Forests" are the collections of "decision trees", which can produce better prediction accuracy. The intention is to create many decision trees based on an subsets of the datasets. At each node of the tree, n number of variables out of the feature sets are selected randomly and the best split on these variables is arrived at.
    - Multiple number of trees can be created on two third of the training data ("roughly 62.3%"). Data is chosen randomly within this sample.
    - Several of the predictor variables from the data are arbitrarily chosen out of all of the predictor variables. In order to split the node, the best possible split from the variables are used. "By default, the amount of the variables selected is the Square Root of the total number of all predictors and it is persistent for all trees".
    - Using the rest of the data from the data set, the misclassification rate is calculated. The total "error rate" gets calculated as the overall "out-of-bag" error rate.
    - Each trained tree provides its own classification result, giving its "own vote". The class that received the most "votes" is selected as the result.
- *Neural Network:* Neural networks have been around for at least 50 years as one of the type of Machine Learning algorithms. The important unit of a neural network is a node, which is roughly based on the similar model like the biological neurons in the human brain. The connections between these neurons are also modelled on biological brains and these connections develop over time ["with training"].

VI. USE CASE

A. Discovery & Attack Surface Assesment

- As connected IoT devices communicate, usually via HTTP initially, it's possible to continuously, automatically mine the involved user agent strings within the traffic, even if the devices communicate over nonstandard TCP ports, in order to discover these active devices and understand their device type, make, model, version, etc.

B. Detection

Continuously analysing the network traffic generated by IoT devices can help to identify potential compromise and misuse [11]. Here are some examples of traffic characteristics that can be monitored for:

- Repeated login attempts to IoT devices, which are using factory default credentials (e.g. admin/admin). This may indicate an attacker or malicious process is attempting to gain access via brute force.
- If IoT devices are moved to their own dedicated network segment / VLAN, monitor for unexpected traffic between the IoT devices and your other internal hosts on the main internal network.
- Outbound SOCKS proxy traffic from IoT devices.
- Communication from IoT devices to public IPs / Domains known to be associated with botnets, command and control, proxy usage, crypto mining, etc.
- The SSL/TLS certificates used to establish secure connections to public servers.
- Dynamically generated domains being looked up from IoT devices.
- Changes in traffic rates and patterns from an established baseline.

As noted earlier, IoT devices must communicate across the network to function and attackers must communicate across the network to repurpose them. With network traffic analysis, security analysts have the means to identify devices of interest and examine the full extent of their activity, even amongst the din of an internet of things.

C. Metrics And Cross Validation

Below four types of commonly used performance indicators are generally used when it comes to detection of malwares:

- "True positive (TP): This indicates that a ransomware is correctly predicted/detected as a malicious application."
- "True negative (TN): This indicates that a goodware is detected/predicted as a non-malicious application correctly."
- "False positive (FP): This indicates that a goodware is mistakenly detected/predicted as a malicious application."
- "False negative (FN): This indicates that a ransomware is not detected/predicted and labelled as a non-malicious application."

The effectiveness of our proposed methods are evaluated by using machine learning performance evaluation metrics which are "Accuracy, Recall, Precision and AUC". *Accuracy* is defined as the number of samples that a classifier can correctly detect, divided by the addition of number of all ransomware and good ware applications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

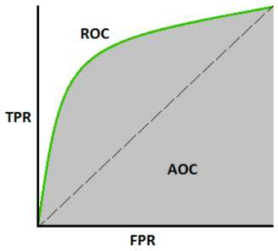
Recall value or the detection rate is the ratio of ransomware samples that are correctly predicted

$$Recall = \frac{TP}{TP + FN}$$

Precision is the calculated ratio of predicted ransomware that are correctly identified as a malware. Precision is defined below

$$\text{Precision} = \frac{TP}{TP + FP}$$

["AUC (Area Under the Curve) represents the probability that a true positive is positioned to the right of a true negative."] AUC ranges in values from 0 to 1. A model which predicts 100% wrong values has an AUC of 0.0 and one which predicts 100% correct values has an AUC of 1.0.



- *Performance of Algorithms*: We have used the leave-one-out technique for cross validation. Below figure illustrate the network traffic usage graph due to ransomware.

D. Data Flow

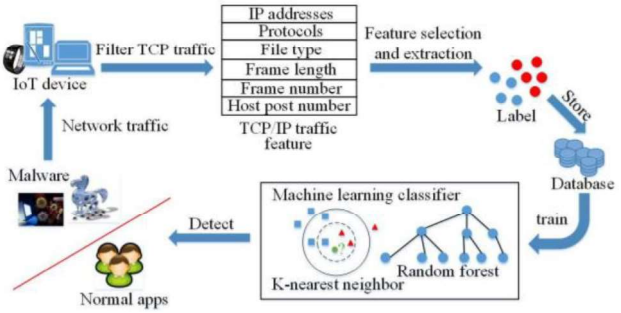


Fig. 14: Data Flow Diagram

(http://www.1proga.info/support-vector-machine-algorithm-389128593a10a109ec0dd4/)

The IoT devices filter various TCP packets, select the required features from various network identifiers including the frame number and the length, labels them and stores them features in a database. The ransomware detection methodology based on KNN assigns the network traffic to the class with largest number of objects among its K nearest neighbours. And the random forest classifier uses the labelled network traffic to build the decision trees to distinguish and identify ransomware. According to the experiments we have done in [14], the accuracy rate of the KNN based detection and random forest-based detection with our dataset are 83.93% and 81.97%, respectively.

- Network Data Set collected from the IOT devices in pcap file format. Below is a sample of the data which was converted and extracted using wireshark from the pcap file format.

No.	Time	Source	Destination Feature	Protocol	Length	Info
1	0	Cisco_db:19:c3	Broadcast	ARP	60	Who has 147.32.84.165? Tell 147.32.84.1
2	8.982709	Cisco_db:19:c3	Broadcast	ARP	60	Who has 147.32.84.165? Tell 147.32.84.1
3	50.099564	Cisco_db:19:c3	Broadcast	ARP	60	Who has 147.32.84.165? Tell 147.32.84.1
4	50.369266	54:52:00:00:00:01	Broadcast	ARP	60	Who has 147.32.84.165? Tell 147.32.84.85
5	51.369054	54:52:00:00:00:01	Broadcast	ARP	60	Who has 147.32.84.165? Tell 147.32.84.85
6	52.369688	54:52:00:00:00:01	Broadcast	ARP	60	Who has 147.32.84.165? Tell 147.32.84.85

The data set in our experiment is is divided into clusters for the model execution and these are called as train and test data sets. As the train and test contain different data sets, they do not show the same attack probability distribution. 16 out of the 38 labelled threats are present with in the test data set in our experiment.

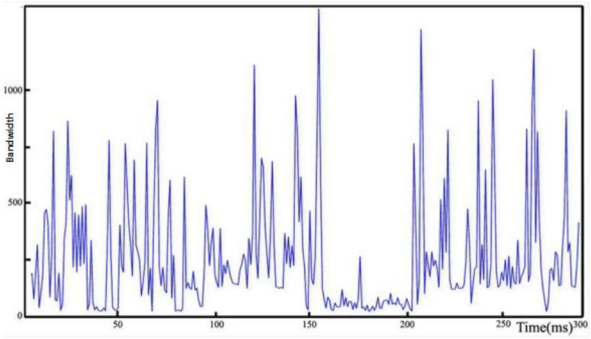


Fig-15 Benign Data

- Network Traffic data analysis on benign data given below. [Fig-15]

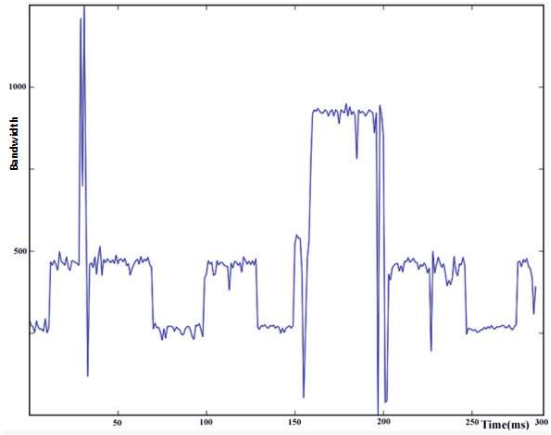


Fig-16 Infected Data with Ransomware

- Network Traffic data analysis on infected data with Ransomware given above. [Fig-16].

Comparison of the above two figures shows there aare significant differences between patterns of network usage for ransomware infected applications versus benign applications.

- *Performance of Machine Learning Techniques: A Comparative Summary*

Model	Accuracy %	Recall %	Precision %	AUC %	F-Measure %
KNN (K = 1)	72.08	72.18	57.6	67.53	63.64
KNN (K = 5)	72.82	71.29	60.48	68.51	64.62
KNN (K = 10)	72.45	72.18	56.1	67.83	63.94
KNN (K = n)	**83.93**	**79.96**	**75.42**	**81.12**	**77.23**
Neural Network	76.16	74.4	63.14	71.79	67.9
Random Forest	81.97	77.74	70.53	77.41	73.52
SVM	78.75	75.51	67.15	74.57	70.68

- A Comparative Summary has been given below explaining the evaluation Metrics for different Window Sizes namely KNN and Euclidean distance:

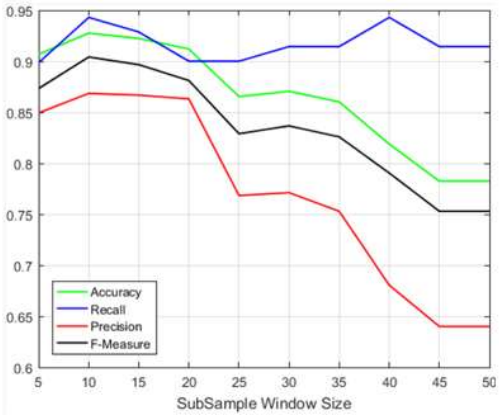


Fig-17

- Evaluation of metrics for different window sizes for KNN and n distance:

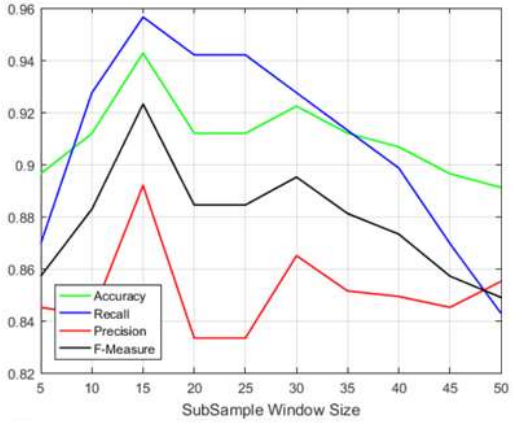


Fig-18

["In KNN's requirement for concurrent distance calculations among training and testing data sets, objects are implemented using parallel processing. Subsamples dictionary can be partitioned into sperate IoT nodes and each subsample is sent to nodes. They return a label and a similarity value and the label having less similarity value is final subsample's label. This approach reduces the classification time and mitigates the need for storage capacity in every node."]

VII. Future Directions of IoT Technologies In Smart City

Smart city technologies and infrastructure are significantly impacted by the contribution of IoT technologies. This has resulted in a large number of business opportunities and wide-ranging growth potential being created. For developing future IoT technologies in smart cities efficiently and successfully, below points require focus and attention.

Important and essential: To have IoT technologies for a smart city infrastructure, we will need "Sensor-oriented" technologies for wireless networking

- Important: While sensor-oriented technologies are quite important ones, however "network services" technologies are considered as the most important IoT technologies for smart city infrastructures.
- Essential: In addition to the above, "energy-related technologies" are also considered as the most essential IoT technologies for the smart city infrastructure. Some of these technologies applied in a smart home environment are important aspects.

Today devices employing IoT technologies need to first have various operating systems which need to include low to high end processors. These can then provide apporiately distributed resources. This can be considered as one of the most important requirement.

Secondly, the currently available data is highly unstructruerd and atypical. This requires data management solutions to be built to handle the massive amounts of data collected by various IoT technology devices. Technologies eed to be developed for data categorization and intelligent analysis.

Third, There is a requirement to develop compatible integrated Sustainability applications which can provide various IoT technology services using appropriate network technologies.

Fourth, there is a need to develop suitable resolutions and planning should done for the purpose of data security and privacy. While users need to connect to services based on IoT technologies, the reliability and privacy in data processing and storing needs to be taken care. We need to have reliable and safe connectivity options from each IoT technology device to the smart city infrastructure.

VIII. Conclusion

With increasing occurrence of Internet-connected devices in our data centric society, securing the IoT networks is critical. IoT nodes that are compromised could hold the network to ransom. For example, blocking availability to data in an IoT network by ransomware could negatively affect the operation of an organisation and may result in significant financial loss in addition to reputation damage.

In this paper, we have presented an approach to detect ransomware, using network traffic flow analysis. We have utilized the distinguished fingerprints of ransomware's network usage pattern to separate ransomware from non-malicious applications. Through our experiments we have demonstrated that our approach could achieve a detection rate of 83.93% and a precision rate of 75.42%.

We have shown that ML is a viable and effective approach to detect new variants and families of ransomware for further analysis, and as a complement to the Anti-Virus solutions. We have shown the approach to be an effective way of automatically selecting the features, while KNN has come out to be an accurate algorithm in case of Ransomware detections. It is easy to train and update, and it's fast.

References

[1] [Online] Available: https://iot-analytics.com/internet-of-things-definition/.

[2] [Online] Available: http://www.gartner.com/newsroom/id/3165317.

[3] Tankard, Colin. "The security issues of the Internet of Things." Computer Fraud & Security 2015, no. 9 (2015): 11-14.

[4] Watson S, Dehghantanha A (2016) Digital forensics: the missing piece of the internet of things promise. Comput Fraud Secur 2016(6):5–8

[5] O'Grman G, McDonald G (2012) Ransomware: a growing menace. Tech. rep., Symantec Corporation. http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/ransomware-a-growing-menace.pdf. Accessed 12 Feb 2017

[6] Caviglione L, Gaggero M, Lalande JF, Mazurczyk W, Urbański M (2016) Seeing the unseen: revealing mobile malware hidden

communications via energy consumption and artificial intelligence. IEEE Trans Inf Forensics Secur 11(4):799–810

[7] Jonas Olsson, Texas Instrument, 6LoWPAN demystified, http://www.ti.com/lit/wp/swry013/swry013.pdf

[8] E. Bertino and N. Islam, "Botnets and internet of things security," Computer, vol. 50, pp. 76–79, 2017.[24] B. Nassi, A. Shamir, and Y. Elovici, "Oops!... i think I scanned a malware," arXiv preprint arXiv:1703.07751, 2017.

[9] R. Richardson and M. North, "Ransomware: Evolution, mitigation and prevention," International Management Re-view, vol. 13, no. 1, p. 10, 2017.

[10] The rise of ransomware and emerging security challenges in the Internet of Things. Available from: https://www.researchgate.net/publication/319527564_The_rise_of_ransomware_and_emerging_security_challenges_in_the_Internet_of_Things [accessed Oct 28, 2018].

[11] [Online] https://www.corvil.com/blog/2018/discovering-iot-devices-and-monitoring-their-network-traffic

[12] Haykin S (1998) Neural networks: a comprehensive foundation, 2nd edn. Prentice Hall, 20(8):2481–2501 Kim H, Smith J, Shin KG (2008) Detecting energy-greedy anomalies and mobile malware variants. In: Proceedings of the 6th international conference on mobile systems, applications, and services. ACM, pp 239–252

[13] Kohavi R et al (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai 14:1137–1145

and mobile malware variants. In: Proceedings of the 6th international conference on mobile systems, applications, and services.

[14] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," Soft Computing, vol. 20, no. 1, pp. 343–357, Jan. 2016.

[15] L. Xiao, Y. Li, X. Huang, and X. J. Du, "Cloud-based malware detection game for mobile devices with offloading,"

IEEE Trans. Mobile Computing, vol. 16, no. 10, pp. 2742–2750, Oct. 2017.