A study of IOT security using Machine Learning approaches.

Avantika Krishnan (20BCY10028,avantika.krishnan2020@vitbhopal.ac.in)

July 2022

1 Introduction

2 Related Works

Syeda Manjia Tahsien et al. [19] discussed the multiple layers of architecture in the Internet Of Things(IoT), connected via various communication protocols which enabled the entire world to be connected. Unfortunately, this made IoT devices vulnerable to a plethora of cyber attacks, with multiple levels of severity and effects. ML techniques like Supervised, unsupervised, and reinforcement learning that are applied to establish a secure environment, were discussed. Francesco Restuccia et al. [15] discussed the unique challenges present in IoT security and the main steps in the process of data collection for ML algorithms. With an increasing number of IoT devices deployed, traditional security is no longer very effective in protecting us from cyber threats. It calls for the use of ML approaches for securing the devices. The paper presented a novel discussion involving systems with security-by-design, where the threats are mitigated using a combination of learning and polymorphic architectures, it also provided us with a survey of existing work on Machine Learning and software-defined Networking for IoT security. We must get to know about the background of large-scale attacks and intrusion detection techniques. Rasheed Ahmad et al. [2] analyzed various research papers from ACM Digital library and IEEE/IET Electronic Library and proposed six key research questions related to the topic at hand. Using these questions, various discussions on topics, including security challenges, impact, preventive measures, and different machine learning techniques such as SVM, RF, and deep learning techniques such as DL, CNN, RNN, and AE used by researchers, were done. Usage of IoT has increased in both private and public sectors. The expectations of the work efficiency from IoT are high, but maintaining it requires smart security protocols. Darko Andročec et al. [5] systematically researched and identified the different types of studies and the primary area of research in using ML-embedded algorithms for IoT security. This paper discussed how it carefully selected research papers which were then analyzed and divided into three categories, intrusion detection,

authentication, and other. If we want to adopt even more widespread use of IoT technology in the future, we need to ensure that we have stronger resistance against attacks. Deep Learning, a part of Machine Learning, is a potential solution in case of security. Unfortunately, there is a lack of systematic review, but Lerina Aversano et al. [6] discussed deep learning specifically by investigating some key-research questions. The analysis also highlighted research gaps that exist. Two major concerns in IoT applications are seecurity and privacy. The aim of using IoT is to connect people and also the work that they do daily. Hence, under this scenario, it is crucial to preserve privacy. Mohammad Amiri-Zarandi et al [4]. examined privacy concerns of IoT devices, which helped group data used by ML-based solutions for privacy protection. Additionally, creative ways were explored where ML-based solutions might use these data sources in the IoT ecosystem. Nazar Waheed et al. [24] discussed the threats and the existing solutions using Machine Learning and Blockchain technology to help us categorize the various requirements under security and privacy. The challenges faced while using Blockchain and Machine Learning, individually and together, were looked into. Wireless Sensor Networks are the primary building blocks of the Internet of Things. Marwa Mamdouh et al. [13] discussed this overlapping research area and classified the different types of attacks, for example, Goal-Oriented and Performance Oriented attacks. Furthermore, discussed how we can counter attacks critical to IoT and WSN, for example, using Neural Network to counter Ddos and Man in the Middle Attacks.

The growth of IoT has allowed the collection of unprecedented data, automation, and smart command into numerous Cyber-Physical Systems. Under such circumstances, it is important to sketch out the problem space of applying Machine Learning techniques. Fan Liang et al. [12] provided an overview of various ML technologies. The benefits and risks of using these dynamic algorithms in Cybersecurity and Cyber-Physical Systems were analyzed under three categories. The categories are The Good, where ML is used for securing IoT devices, the Bad, where ML algorithms get deceived by the attacks and the Ugly, where ML is used for implementing attacks. IoT is a combination of Internet devices and intelligent communication devices. Advancements in technology have increased the use of IoT devices in our daily lives. Unfortunately, the interconnectedness of devices increased their vulnerability. Ankit Thakkar et al. [21] provided a comprehensive study of various techniques and solutions used and the multiple security challenges that exist. The recent updates in this domain were also discussed. Currently, research on finding the most efficient machine learning algorithm for protecting IoT devices against Dos attacks is a popular area of study. Abhishek Verma et al. [23] presented discussions on classification algorithms and assessed the performance of classifiers such as random forests, AdaBoost, gradient boosted machines, etc, using popular datasets. The experiment showed that extreme gradient boost and classification and regression trees are the best choices for intrusion detection in IoT devices. Furthermore, Raspberry Pi was used to evaluate the average time response.

To help fight the rising problem of dynamic attack platforms, José Roldána et al. [16] proposed a framework combining Complex Event Processing (CEP)

and Machine Learning. CEP can efficiently handle network security without previously storing data, a technology seldom used by systems. This proposed framework is an extension of Event-Driven Service Oriented Architecture (SOA) platforms. It was tested against a healthcare network to check its accuracy, and the results were satisfactory. Ayush Kumar et al. [10] proposed a modular solution for intrusion detection consisting of a packet traffic feature. A database stored the traffic patterns of malware, which can be updated. If the patterns match, the given policies decide what needs to be done next. The performance was analyzed using testbed experiments which revealed that if benign and malicious traffic has significant differences, then the model can easily classify them. Otherwise, it won't give accurate results. Hence, more research is required so we can help model an ideal solution. Softwarized networks are another possible area of study. Miloud Bagaa et al. [7] discussed a framework where it combined two different paradigms, Software-defined networking (SDN) and Network Function Virtualization (NFV), to improve security. The framework contained two main layers Security Orchestration Plane and Security Enforcement Plane. SDN contributes to network mechanization and NFV in services together, they can meet a wide range of requirements for cyber attacks.

Internet of things combined with mobile computing devices is a new approach coined with the term mobile Internet of things. Igor Konteko et al. [9] discussed a novel framework merging the concepts of Machine Learning and Big data processing for security monitoring of the mobile internet of things. The architecture design was based on two modes, training mode, and analysis mode. The classifier operation results were subject to plurality voting, weighted voting, and soft voting. It also offered discussions containing mathematical foundations and a thorough experimental analysis of the framework. Machine Learning uses data sets to detect intrusion, but this creates a pathway for yet another form of cyber attack called poisoning attacks, where the data gets manipulated to generate complications. Nathalie Baracaldo et al. [8] proposed a framework that uses provenance or meta-data, which two of the previous works, Reject on Negative Impact (RONI) and the Probability of Sufficiency (PS), do not use. The results of RONI were considered as a baseline, and the proposed model surpassed those performances. Still, we require more research because if the dataset has more than 25% manipulation, then the architecture failed to work. Zolanvari et al. [25] discussed using an experimental setup with a testbed resembling an industrial plant and analyzed areas where ML falls short. The paper focused on analyzing the effects of imbalanced datasets, specifically in the Industrial Internt of Things (IIoT). An artificial neural network was tested against different ratios of imbalanced datasets. Multiple metrics like accuracy, False Alarm Rate(FAR), and Undetected (UR) Rate were used. The paper showed the extent to which Machine Learning algorithms are useful. Along with discussing how ML algorithms can be a potential solution, we need to discuss which algorithm will be most suitable and efficient.

Denial of Service Attacks or Distributed Denial of Service attacks is a common and very disruptive area of concern for IoT devices. Furthermore, when combined with Botnets they are more disruptive than they are, which is a catas-

trophe hence, Reem Alhajri et al. [3] discussed using auto-encoders as a potential solution to detect IoT botnets. The most concerning point to mention here is the constant mutation of intricate infrastructures of such attacks. With the explosion of IoT devices in the healthcare industry scholars have been focusing on using bio-signals as secret keys for the encryption of data. Sandeep Pirbhulal et al. [14] analyzed the various security laws and requirements for E-healthcare, such as Data integrity, privacy, Data Freshness, etc, and also proposed a framework that utilizes Electrocardiogram (ECG) signals for generating random and unique keys. Muhammad Shafiq et al. [17] proposed a framework and discussed how to select the best ML algorithm. The framework used the Bijective soft Set, a mathematical technique used for decision making. The results are analyzed using five parameters: accuracy, precision, recall, true positive rate, and time taken to build a model.

To summarize and expand our understanding further, a table was created with 8 parameters. This was to analyse the research papers in depth. The parameters included Systematic Review, In-depth discussion on ML techniques, Proposed a novel framework, Performed Technical Analysis, Discussion/Assessment of False Positive Rate, Discussion/Assessment of True Positive Rate, Discussion/Assessment of Datasets, and Future Works.

AUTHOR, YEAR	KEY CONTRIBUTION	Systematic Review	In-depth discussion on ML techniques	Proposed a novel framework	Performed Technical Analysis	Discussion/Assessment of False Positive Rate	Discussion/Assessment of True Positive Rate	Discussion/Assessment of Datasets	Future Works
Syeda Manjia Tahsien et al. , 2020	Discusses the basic concepts of IoT security using ML embedded Algorithms	N	Y	N	N	N	N	N	Y
Francesco Restuccia et al. , 2018	Provided a roadmap for future researchers	N	N	N	N	N	N	Y	Y
Miloud Bagaa et al. , 2020	Proposed a framework combining SDN and NFV	N	Y	Y	Y	Y	Y	Y	Y
Rasheed Ahmad et al. , 2021	Conduct a systematic review for future researchers	Y	Y	N	N	Y	Y	Y	Y
Sandeep Pirbhulal et al. 2019	Present a ML based biometric security approach	N	N	Y	N	N	N	N	Y
José Roldána et al, 2020	Proposed and architecture combining CEP and ML	N	Y	Y	Y	Y	Y	Y	Y
Fan Liang et al. , 2019	Assessed the positive and negatives of ML	N	Y	N	N	N	N	Y	Y
Igor Konteko et al. 2018	Proposed a framework combining Big Data processing and ML	N	Y	Y	Y	Y	Y	Y	Y
Darko Andročec et al. 2018	Create a starting point for further research on the topic	Y	Y	N	N	N	N	N	Y
Abhishek Verma et al. 2020	Analyzed performances of seven classifiers	N	Y	N	Y	Y	Y	Y	Y
Mohammad Amiri-Zarandi et al 2020	Analyzed privacy in IoT including topics like scalability, interoperability and resource limitation.	N	Y	N	N	N	N	N	Y
Nazar Waheed et al. 2020	Discussed both ML and BC techniques for security in IoT	Y	N	N	N	Y	Y	Y	Y
Ayush Kumar et al. 2019	Proposed a modular solution for IDS	N	N	Y	Y	Y	Y	Y	Y
Maede Zolanvari et al. 2018	Analyzed how ML algorithms works with imbalanced data sets	N	Y	N	Y	Y	Y	Y	N
Reem Alhajri et al. 2019	Analyzed feasibility of using Aut-encoders against Ddos Attacks	N	N	N	N	N	N	N	Y
Ankit Thakkar et al. 2021	Presented a comprehensive study analyzing risks,challenges, and updates	N	Y	N	N	Y	Y	Y	Y
Lerina Aversano et al. 2021	Analyzed landscape of Deep Learning in IoT and identified research gaps	Y	Y	N	N	N	N	Y	Y
Marwa Mamdouh et al. 2018	Investigate threats that affects WSN and IoT	N	Y	N	N	N	N	N	N
Muhammad Shafiq et al. 2020	Proposed a framework for selecting the most efficient ML algorithm	N	Y	Y	Y	N	Y	Y	N
Nathalie Baracaldo et al. 2018	Proposed a model to identify poisonous attacks	N	N	Y	Y	Y	N	Y	N

3 Methodology

3.1 BACKGROUND

To know any topic, we need to have a thorough understanding of its keywords. Hence, Let us understand the basic terminologies of our topic.

3.1.1:WHAT IS IoT

The Internet of Things can be described as a network of physical elements. These could be mechanical, digital, and even biological like humans or animals. In this network, each node is represented by any 'thing' that can be assigned an Internet protocol address and can communicate with other nodes. These could be a person with a heart monitor, dogs with digital collars to track them, air conditioners and refrigerators with Bluetooth devices, or auto-mobiles with sensors to alert the drivers if they are falling asleep.

3.1.2 : HOW IoT WORKS?

IoT works with data that has been collected from its environment. This data is then sent to the cloud using methods like cellular, satellite, WiFi, Bluetooth, etc. The method will depend on the requirements of the device. The data will then be processed which can either be simple or complex according to the demands of the device. Simple could include things like checking temperatures and complex can include things like identifying movement. After the data has been collected and processed, action is taken according to the results obtained. The results might be sent to the user via an email, it could also alert the user under the circumstances that an abnormal or unexpected event has occurred. For example, if movement is detected in an empty house.

3.1.3: WHAT IS MACHINE LEARNING

Machine learning is a technique used to imitate the way humans learn, behave and react. It is an upcoming domain in the realm of Artificial Intelligence and computer science. In very simple terms using Machine Learning, we can give systems the ability to think. This is a very innovative technology with tons of futuristic applications. It uses various mechanisms to implement actions including giving us personalized Netflix recommendations, keeping self-driving cars safe from accidents, observing customer trends to improve marketing, etc.

3.1.4: HOW MACHINE LEARNING WORKS

If we explain machine learning in very simple terms it takes datasets, processes them, and gives us results. But there are much more complex systems at play as with each scenario the algorithm learns and keeps increasing its accuracy. UC Berkeley divides the machine learning processes into three parts. One is Decision Process, where the algorithm, based on some input data, gives us results or a prediction. Two, Error Function: error functions serve as a comparison parameter to improve the accuracy of the model. Three: Model Optimization Process: A known example is used to constantly balance, and update the mechanisms for constant optimization.

3.1.5: CATEGORIES OF MACHINE LEARNING

There are three categories of Machine Learning.
1.Supervised learning: This method gives results by processing labeled datasets. The model goes through constant optimization to accurately fit a dataset. It is made sure that overfitting and underfitting are avoided. Supervised learning is most popularly used in BioInformatics where it stores the information of humans like fingerprints, DNA, iris texture, etc. Based on whether the datasets are discrete or numerical we have the following categories:

- **Classification:** The output here is a discrete value. This means there are a set of results like [True, False] and the result will be either one of them. A few mechanisms which implement classification learning are Support Vector, Random Forest, Bayesian Theorem, etc.
- **Regression:** The output here is an integer value and can also be continuous. For example, we can say that the file has a 50% chance of being malware. A few mechanisms that implement Regression learning are Neural networks, Decision Tree, etc.

2.Unsupervised Learning: The method processes unlabeled data, and there is no output generated. This method is used to simply compare and categorize the inputs into clusters. For example, an email can be categorized as spam using this method. K-means Clustering technique and Principal Component Analysis are some methods used to implement Unsupervised learning.

3.Reinforcement Learning: It is an intermediate between supervised and unsupervised learning. It has no labeled datasets but a reward output. It learns from its environment, constantly making changes to maximize performance. There is no predefined set of actions and is dependent on the trial and error method. Two important reinforcement methods are Policy search and value function approximation.

3.1.6: SECURITY OF IoT USING MACHINE LEARNING

IoT security has become a major issue for IT teams. IoT devices are used in every sector for example government and private businesses, healthcare, agriculture, banks, etc. Machine Learning has the potential to address some of the unique challenges that IoT has. Even though IoT is the weakest link in a network it is still used widely due to its property of scalability. It helps in the easy addition of new users in every sector. At the most, basic level ML can help in identifying every new node in the network. Using automated scanning can keep track of all the devices and also keep them safe. Ml mechanisms can be trained to resist attacks by identifying irregular patterns and behavior that could be a potential attack. ML is designed to keep up with the changing threat platforms. This means that ML is an ideal solution against Zero-day threats, a very urgent

area of concern in cybersecurity.

3.2 PROBLEM STATEMENT

In using ML for IoT security we have two revolutionary approaches which utilize physical layer authentication for access control. These are game theoretic approaches and machine learning techniques. They will help in differentiating between malicious and normal traffic. Machine learning has to be such a platform that learns on its own and customizes itself to identify irregular patterns, which is why Reinforcement learning is being focused on building security mechanisms that can tackle zero-day attacks. As adversaries keep finding new loopholes every day, the algorithms need to be trained to identify them. Miloud Bagaa et al. [7] and José Roldána et al, [16] proposed novel solutions where the former combines SDN, which helps in connecting with the hardware and introduces dynamicity, NFV, which virtualizes the dedicated hardware and AI, while the latter combines Complex event processing along with ML. Another major issue is the datasets that are being fed into the mechanism. They can be manipulated to create unexpected results. To defend against poisoning attacks Nathalie Baracaldo et al. [8] presented a novel framework to detect them using the concept of segmentation. Imbalanced datasets are also an issue which is why Maede Zolanvari et al. [25] in his paper tested the limits to which ML is useful in cases where we have an imbalanced dataset. Selecting the right ML algorithm is equally important and it requires a thorough analysis. Muhammad Shafiq et al. [17] proposed a method involving a bijective soft approach for the same purpose. A topic seldom discussed is how the incredible data analysis capabilities of ML can be used for attacks that Fan Liang et al.[12] have analyzed.

The main contribution of this paper is to review the work done by multiple researchers to enable us to understand the areas of concern that need to be looked into. This will help future researchers quickly understand the concepts that need to be prioritized to make security in IoT stronger. Furthermore, possible areas of future research have been provided with a structure of architecture that might benefit researchers. Detailed analysis on using Machine Learning as a security mechanism for IoT was conducted. The first papers were collected and this included both technical and survey-oriented papers to truly understand the level of work done in this domain. The study was then presented under a few parameters in a table. We can see that authors Miloud Bagaa et al. [7], Sandeep Pirbhulal et al. [14], José Roldána et al, [16], Igor Konteko et al. [9], Ayush Kumar et al. [10], Muhammad Shafiq et al. [17] and Nathalie Baracaldo et al. [8] have given a novel framework, which acts as a solution on how to use ML for security. This will help future researchers understand and improve upon these frameworks and increase their efficiency. While Abhishek Verma et al. [23], and Maede Zolanvari et al. [25], have studied the effects of certain datasets and algorithms using the parameter of technical analysis. The parameters False positive rate and True positive rate are important areas of concern and authors like Rasheed Ahmad et al. [2], Nazar Waheed et al. [24], and Ankit Thakkar et al. [21] even after having no technical aspect have provided some discussion

upon it. Although datasets have been discussed by multiple authors we need to further look into how these can be used to exploit the defense mechanisms, a topic discussed in papers written by Nathalie Baracaldo et al. [8] and Maede Zolanvari et al. [25]. We need more research in this area where rather than focusing on ML we focus on the datasets as they are processed by the algorithms and can be easily manipulated.

3.3 Discussion of existing solutions

Multiple researchers have tried addressing the domain of using Machine Learning for strengthening IoT security. It has been done in many ways. Roadmaps for future works have been provided using thorough research by researchers like Syeda Manjia Tahsien et al [19], Rasheed Ahmad, et al. [2], and Darko Andročec et al. [5]. Frameworks have been provided which can be used as a basis for building architectures. Furthermore, researchers have also provided us with various solutions which include architecture and analysis. All these architectures have their combination of various modules but the basic framework beneath remains the same. The following figure tries to portray a general architecture, that tries to include a skeleton structure of what most researchers have discussed and tried to build upon.

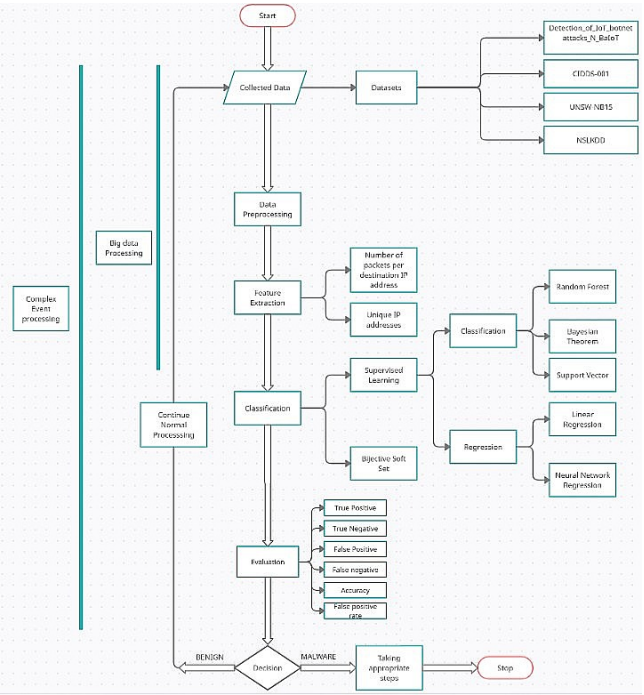

Figure 1: General Architecture

In fig 1. We can see the architecture of how an ML can be used for IoT

security. The explanation for each component is given below -
Collected Data - The traffic generated due to the network of IoT devices acts as data. Machine learning algorithms work on this data and generate results. For technical analysis researchers sometimes use their testbeds to generate data as done by Maede Zolanvari et al. [25] by using an industrial testbed that included sensors. To make research work easier we also have predefined datasets such as -
DetectionofIoTbotnetattacksNBaIoT - This contains the network traffic information between mobile devices. To be precise it utilizes 9 devices. It stores information as a CSV file and contains 11 classes, 10 for attacks and 1 for benign. It also contains many duplicate values.
CIDDS-001 - Containing approximately 32 million records it is a recent dataset generated for research and development in the area of network intrusion detection. It was created using python language scripts.
UNSW-NB15 - It contains information regarding 9 attacks and has been divided into two parts: train and test. Both sets have instances of normal traffic and attack traffic. Normal traffic instances in the train set are 56,000 while in the test set it is 37,000. Attack traffic instances in the train set are 119,341 and in the test set, it is 45,332.
NSLKDD - It offers a platform for comparing the performance of different IDS methods. It helps in validating the performances of classifiers. It also contains training and test sets consisting of attack and normal traffic. Normal traffic instances in the train set are 11,743 while in the test set it is 12,833. Attack traffic instances in the train set are 13,499 and in the test set it is 9,711

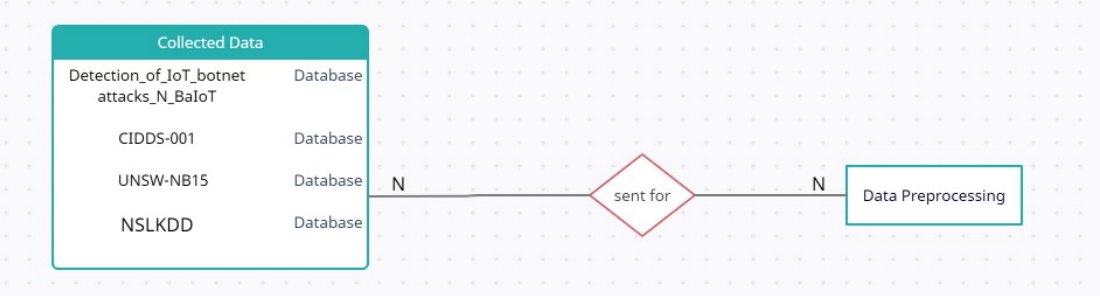

Figure 2: Relationship between collected Data and Data Preprocessing

2. Data Preprocessing - The data collected cannot be directly sent for feature extraction or into the ML algorithms. The raw data needs to be first preprocessed. It is done to make it suitable for the ML models which will be used. This will help in increasing the accuracy of the performance. For example, in the dataset "DetectionofIoTbotnetattacksNBaIoT" there are duplicate values that need to be removed. Hence, through preprocessing the duplicate values are removed to reduce load.

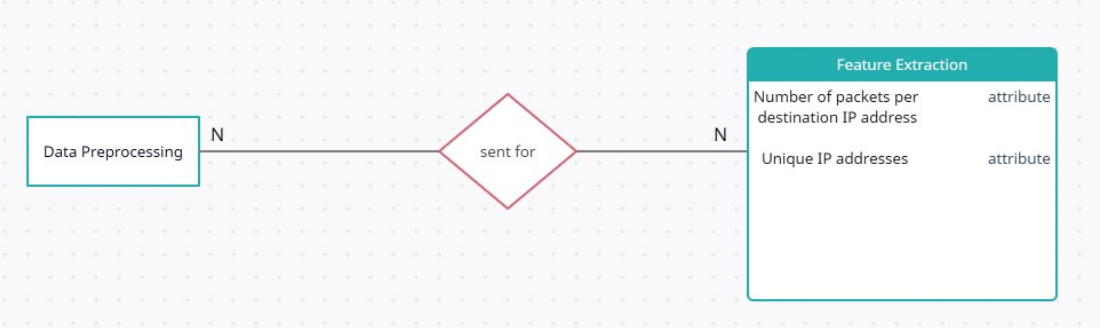

Figure 3: Relationship between Data Preprocessing and Feature Extraction

3. Feature Extraction - Under this, we classify the traffic under certain features according to the requirements of the ML model. The ML model can then analyze it and classify them. Ayush Kumar et al. [10] took the following features for classification -

Number of packets per destination IP address - Certain flooding attacks can be identified by knowing the maximum number of packets per destination. Hence, extracting this feature will help us identify those.

Unique IP addresses - This feature is extracted during certain malware attacks the number of unique and probably unknown IP addresses trying to connect with a device might increase hence we need to analyze that.

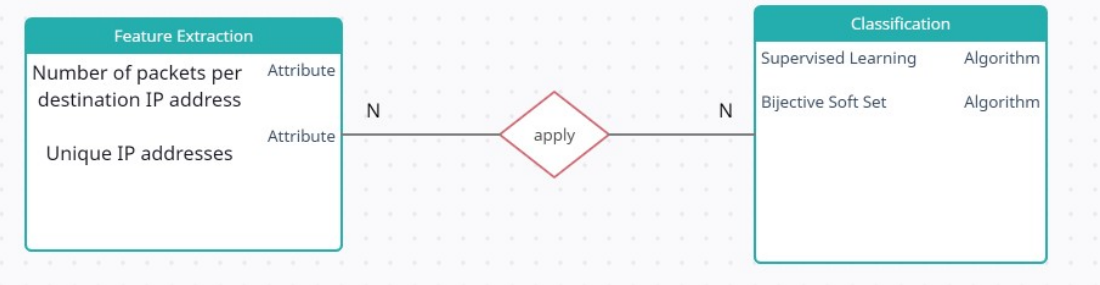

Figure 4: Relationship between Feature Extraction and Classification

4. Classification - Here we arrange the data into different categories or groups. This broader aspect can help us differentiate between Benign and malware-induced traffic.

Supervised Learning - As discussed in 3.1.5 Supervised Learning uses labeled datasets and has 2 categories -

1. Classification - The output here is discrete. Some common algorithms are -

- **Support Vector Machine** - A hyperplane helps separate sets of data into two parts. Hence, there are multiple ways a hyperplane could be drawn. This algorithm tries to define the most suitable hyperplane such that the distance between two data points of both classes is maximum.

- **Random Forest** - It contains a number of decision trees that will contain subsets of the collected data. The average of all the predictions is taken to ultimately give us the result. Its accuracy is very high and the time taken is low. The more the number of Decision trees, the more will be accuracy.

- **Bayesian Theorem** - It is based on the Bayes Theorem. In this method, the algorithm assumes independence of the data from each other.

2. Regression Learning - As discussed in 3.1.5 in Regression Learning the output is an integer value and is continuous. It has 2 categories -

- **Linear Regression** - It establishes a relationship between an independent and a dependent variable with the help of a straight line. It helps us predict unknown values outside of the data points that are not on the line but can assume will exist on the line.

- **Neural Network Regression** - It has the ability to learn the complex relationships between attributes that are non-linear. The output will be based on a function of the inputs.

  There are other classifiers as well like unsupervised learning and reinforcement learning which have been discussed before but usually, they are not used for testing purposes as they are hard to work with. They do not contain labeled data like supervised learning which is much more efficient to test models.

Bijective Soft Set - It is a mathematical tool to select the most effective classifier for a given scenario. Based on the soft set approach it utilizes intersection and union operations on the rows and columns for greater accuracy of results. A soft set involves the concept of a universal set and a parameter set that has its attributes. In the research community, this method has proved to be very useful and hence it was also adopted by Muhammad Shafiq et al. [17] for effective selection of ML algorithms.

Figure 5: Relationship between Classifiers and Evaluation

4. Evaluation: We evaluate the different groups and results of the ML classifiers, to draw meaning from them. This helps make a decision as explained ahead. Researchers also use evaluation metrics which is a common way that researchers further evaluate the performances of the model to understand how accurate the results are. Certain common metrics used are -
True Positive (TP) - It indicates the correctly identified malware or threat classes

True Negative (TN) - It indicates the correctly identified absence of malware or threat classes
False Positive (FP) - It falsely indicates the presence of malware or threat class

False Negative (FN) - It falsely indicates the absence of a malware or threat class

Accuracy - It indicates the proportion out of the total results that have been correctly identified. The formula for the same is -
TP + TN / TP + TN + FP + FN

False Positive rate (FPR) - It is the proportion of falsely indicated presence of malware or threat class. The formula for the same is -
FP / TN + FP

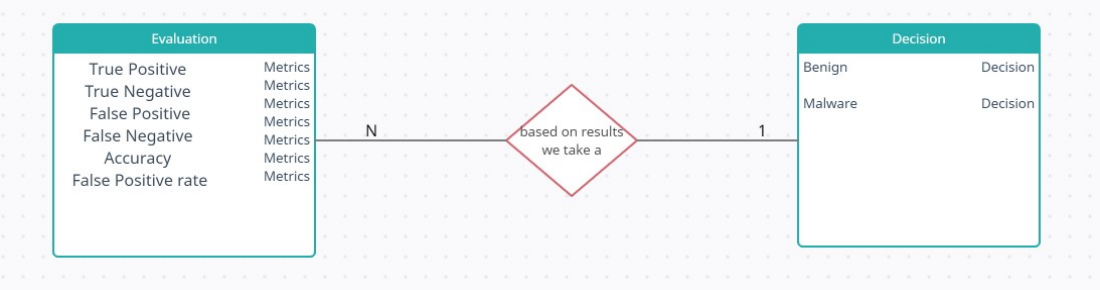

Figure 6: Relationship between Evaluation and Decision

5.Decision: After evaluation we, can make a decision. We can stop the traffic , send an alert to the proper users or authority who can handle the situation more appropriately if malware has been encountered, or if the system is safe we can tell it to continue its normal working.

6. Big Data Processing: It works on very large data which can be structured or unstructured. It extracts meaning from these big data for analysis. It includes steps like Data extraction and modification, data loading, data visualization, and applying ML algorithms to them. It also can be included in data collection, preprocessing, and feature extraction.

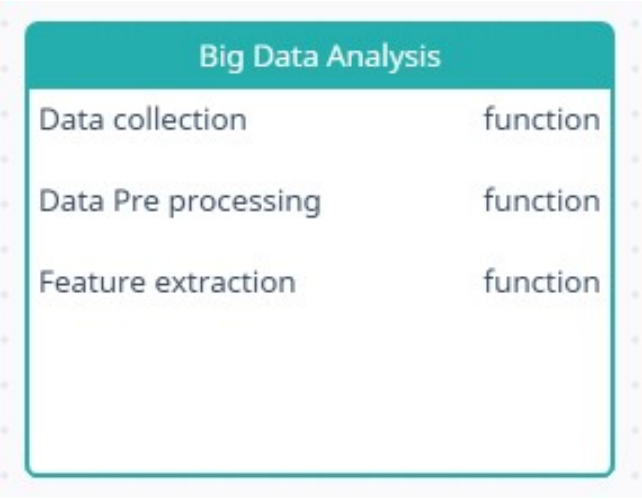

Figure 7: Big Data Processing

7. Complex Event Processing - It is a technology that helps us capture, process, and analyze events. It includes all data collection, feature extraction, classification, and decision. It queries the data collected after matching it to predefined patterns. The incoming traffic or raw data which is collected through IoT devices is processed, and the relevant situations are then highlighted. Every CEP has its own Event processing Language (EPL) which defines patterns.

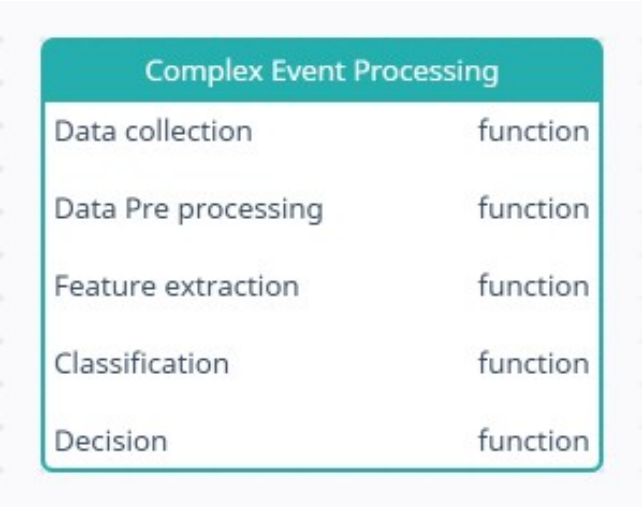

Figure 8: Complex Event Processing

In a general setup for evaluation purposes like implementing frameworks or performing analysis, researchers use various tools and set up testbeds to implement their test cases. For example, Miloud Bagaa et al. [7] in their paper used ONOS SDN CONTROLLER and ETSI OPEN SOURCE MANO (OSM). While José Roldána et al, [16] has used something called MEdit4CEP. They then inject test cases or databases which have attributes.

- **ONOS SDN CONTROLLER** - Open Network Operating System is an open source built to create SDN operations written in Java language. It is developed in a way that it has excellent availability, high horizontal scalability, and high performance.
- **ETSI OPEN SOURCE MANO (OSM)** - It is an open source community-led project developed for NFV Management and Orchestration. It helps in deploying NFV-based operations according to user requirements. Its main parts are Service Orchestrator, Resource Orchestrator, and VNF configuration and Abstraction.
- **MEdit4CEP**. - It is a system used for processing heterogeneous data. It combines stream processing and Complex Event Processing. It can graphically edit CEP and event pattern definitions. It also provides code generation and deployment.

We have multiple open source platforms also which provide user-friendly tools. These usually require low coding experience and some knowledge about ML. Mostly used for testing models. For example -

- **Jupyter Notebooks** - One of the most used platforms to run ML codes. It can support 3 languages: Julia, Python, and R. The name Jupyter has been rightly derived from the combination of the names of these 3 languages. It is used most commonly with Python codes. It is a shared platform where codes can be shared live and also accessed using various GUI.

- **TensorFlow** - Usually used for large ML codes, it is an open-source Python-friendly platform. The data that is used here can include images, audio, and or text files which are far more complicated to work with. Along with machine learning, it also includes neural networks.

- **Amazon Machine Learning (AML)** - It is an open-source cloud-based platform that is used to build and test machine learning models. It reads data from multiple resources like redshift, RDS, and Amazon simple service storage.

The above tools have been summarized in the table below.

Name	Languages Supported	Open Source	Reason Used for	Additional Support
ONON SDN CONTROLLER	JAVA	Y	Create SDN Operations	OpenFlow
ETSI OPEN SOURCE MANO (OSM)	Python 3	Y	Create NFV Operatetions	-
MEdit4CEP.	Not required, automatically generated	Y	Allow grpahical editing of event patterns	PCPN
Jupyter Notebooks	Julia, Python, and R	Y	Running ML codes	Python Libraries
TensorFlow	Python 3.7 to 3.10	Y	Running arge ,complicated ML codes	Decision Forests
Amazon Machine Learning (AML)	-	Y	Build and Test models	Redshift, RDS, and Amazon simple service storage .

As mentioned earlier testbeds are used to test the working of frameworks. These test beds can be pre-developed and just configured appropriately to make them suitable for current scenarios or could be developed from scratch by the researchers. For example, Ayush Kumar et al. [10] used a testbed that used multiple IoT devices to collect incoming and outgoing data. Raspberry Pi is a common platform to perform analysis. Abhishek Verma et al. [23] used it to know the average response time of the classifiers. Maede Zolanvari et al. [25] used a testbed that resembled an industrial internet of things. When researchers use real-world testbeds it makes the results more realistic and helps in solving real-world problems. The IIoT testbed involved a water storage system and data involved were related to it like alarms, logic controllers, turbidity sensors, etc. Hence, to summarize a typical setup will involve data, tools, testbeds, and parameters for results.

4 Conclusion and future work

Based on studying and discussing the various aspects of security in IoT there are certain techniques, areas, or trends where further research might or should occur.

- **Requirement of Dynamic Access Control** - Instead of IoT devices allowing users or devices into the network based on permission and databases, systems should be able to perform real-time scans, of things such as social interactions, etc, to identify the level of access needed to be given. This additional data can further be used to enhance privacy and security.

- **Requirement of Hybrid Models** - Each framework provided by researchers has its advantages and disadvantages. They have their percentage of positive and negative performances. Hence, I feel we are required to combine such frameworks to provide an optimum solution. We need to combine them in such a way that we decrease the negatives and increase the positives.

- **The requirement to pay more attention to datasets** - Researchers most of the time pay most of their attention to ML algorithms and how to improve frameworks involving them. One thing which has been seldom looked at is how datasets can be manipulated to bypass security. Datasets could have enhancements or undocumented code segments which can bypass security, produce unexpected results to crash the system, or deploy malware. Hence, research is required where we use ML or other means to make sure the datasets are benign and true before allowing them to enter the classifiers.

- **Requirement of using cloud services** - Concepts like Machine learning and Deep Learning are resource hungry. To acquire accurate results proper data needs to be fed in. Also nowadays there are attempts to

make everything dynamic which means there has to be a constant collection of data. Zero-day attacks require us to be dynamic as these are attacks that exploit undiscovered security loopholes. Also to make these security services scalable we not only need tons of data but also space for more data. Hence, resource expensive workload should be shifted onto the cloud. Organizations should use cloud/edge services to perform expensive computations to efficiently use space and resources.

Based on the above discussion we can see that in the future there might be new components included in the generic frameworks as in figure 1. If we try to make a framework to include some of the above points then it may look the following.

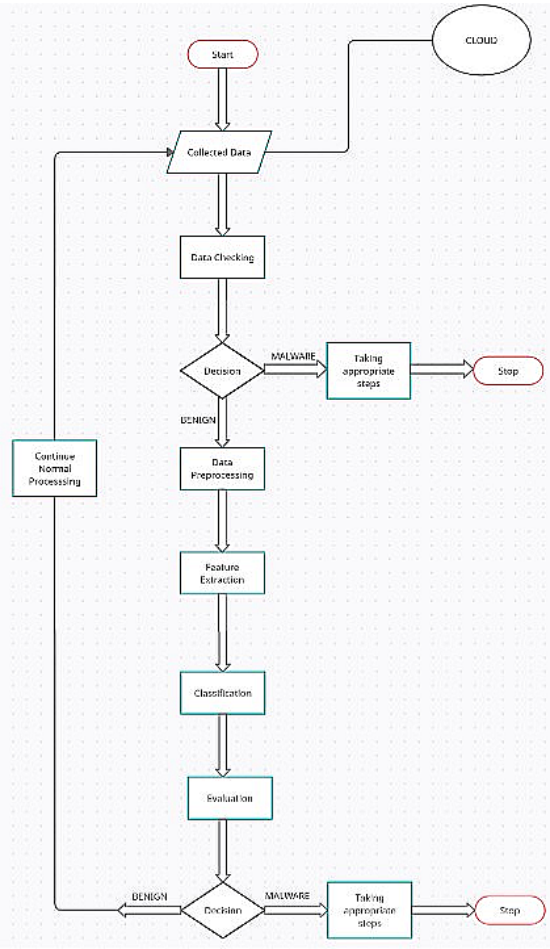

Figure 9: Future Architecture possibility

As discussed in section 3.3 the basic components like Data collected, Data preprocessing, feature extraction, classification, Evaluation, and Decision remain the same. A new addition which we can observe from the above diagram is Cloud services and Data Checking

- **Cloud** - It will play an important role. Not only the majority of data will be stored there, but perhaps certain complex computations will also be performed to reduce load.
- **Data Checking** - Before sending the resources further into the system they should be analyzed to see if it contains any unwanted code snippets or erroneous data which can result in the system crashing. This I feel is important as there are many ways hackers can find to manipulate this data to benefit them. As ML algorithms learn through patterns and ob-

servations, such falsified data can teach them to function in a way that can have very dangerous effects on the security and privacy of a system, potentially allowing cyber attacks to infiltrate.

To infer the above points in depth research had been done on the multiple papers which collected both survey type and technical. Below is a diagrammatic representation of the entire process.

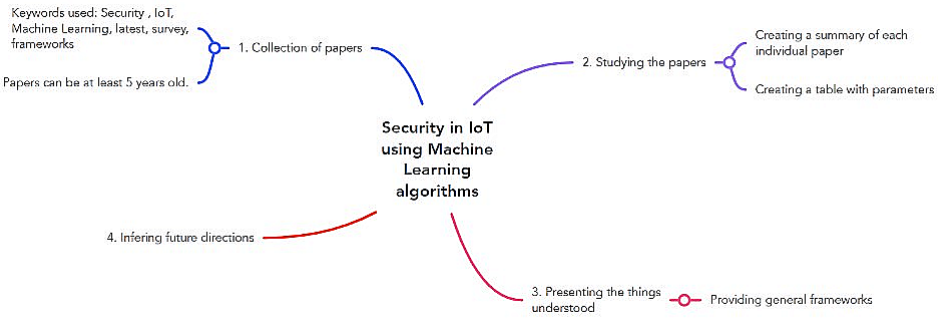

Figure 10: Work Summary

- **Collection of papers** - The papers were collected from reputed organizations like IEEE. Keywords like Security, IoT, and Machine Learning were used. The papers could be only 5 years old or less. Hence, papers from the year 2018 have been considered. Anything before that has been ignored.
- **Studying the papers** - The papers were read and a summary of each was created for easier comprehension. The summary included positive points, negative points, outcomes, and goals. A table was then created with parameters that were deemed important and relevant to the current topic.
- **Presenting the things understood** - After reading the papers a general understanding, problem statement, and discussion have been provided for readers and researchers to know more about this domain. A general framework structure and a discussion about its components have also been given
- **Inferring future directions** - After a thorough discussion and study, a few points were inferred which could be possible future directions for further research.

In this paper, we have tried to broadly understand the steps taken for Security and privacy in the IoT platform. We have studied the frameworks and research

done by eminent scholars and presented our inferences. The main aim was to understand the areas of future research which will improve the security of IoT devices. As we continue to see growth in this area the demands keep increasing and become more complex hence, constant research to keep up with the trends is a necessity.

References

[1] Machine learning tools - javatpoint.

[2] R. Ahmad and I. Alsmadi. Machine learning approaches to iot security: A systematic literature review. *Internet of Things*, 14:100365, 2021.

[3] R. Alhajri, R. Zagrouba, and F. Al-Haidari. Survey for anomaly detection of iot botnets using machine learning auto-encoders. *Int. J. Appl. Eng. Res*, 14(10):2417–2421, 2019.

[4] M. Amiri-Zarandi, R. A. Dara, and E. Fraser. A survey of machine learning-based solutions to protect privacy in the internet of things. *Computers & Security*, 96:101921, 2020.

[5] D. Andročec and N. Vrček. Machine learning for the internet of things security: a systematic. In *13th International Conference on Software Technologies*, volume 4120, page 97060, 2018.

[6] L. Aversano, M. L. Bernardi, M. Cimitile, and R. Pecori. A systematic review on deep learning approaches for iot security. *Computer Science Review*, 40:100389, 2021.

[7] M. Bagaa, T. Taleb, J. B. Bernabe, and A. Skarmeta. A machine learning security framework for iot systems. *IEEE Access*, 8:114066–114077, 2020.

[8] N. Baracaldo, B. Chen, H. Ludwig, A. Safavi, and R. Zhang. Detecting poisoning attacks on machine learning in iot environments. In *2018 IEEE international congress on internet of things (ICIOT)*, pages 57–64. IEEE, 2018.

[9] I. Kotenko, I. Saenko, and A. Branitskiy. Framework for mobile internet of things security monitoring based on big data processing and machine learning. *IEEE Access*, 6:72714–72723, 2018.

[10] A. Kumar and T. J. Lim. Edima: Early detection of iot malware network activity using machine learning techniques. In *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, pages 289–294. IEEE, 2019.

[11] N. Kumar. Top emerging machine learning trends for 2022, Mar 2022.

[12] F. Liang, W. G. Hatcher, W. Liao, W. Gao, and W. Yu. Machine learning for security and the internet of things: the good, the bad, and the ugly. *IEEE Access*, 7:158126–158147, 2019.

[13] M. Mamdouh, M. A. Elrukhsi, and A. Khattab. Securing the internet of things and wireless sensor networks via machine learning: A survey. In *2018 International Conference on Computer and Applications (ICCA)*, pages 215–218. IEEE, 2018.

[14] S. Pirbhulal, N. Pombo, V. Felizardo, N. Garcia, A. H. Sodhro, and S. C. Mukhopadhyay. Towards machine learning enabled security framework for iot-based healthcare. In *2019 13th International Conference on Sensing Technology (ICST)*, pages 1–6. IEEE, 2019.

[15] F. Restuccia, S. D'Oro, and T. Melodia. Securing the internet of things in the age of machine learning and software-defined networking. *IEEE Internet of Things Journal*, 5(6):4829–4842, 2018.

[16] J. Roldán, J. Boubeta-Puig, J. L. Martínez, and G. Ortiz. Integrating complex event processing and machine learning: An intelligent architecture for detecting iot security attacks. *Expert Systems with Applications*, 149:113251, 2020.

[17] M. Shafiq, Z. Tian, Y. Sun, X. Du, and M. Guizani. Selection of effective machine learning algorithm and bot-iot attacks traffic identification for internet of things in smart city. *Future Generation Computer Systems*, 107:433–442, 2020.

[18] J. Shalamanov. 2022's most relevant machine learning trends, Sep 2022.

[19] S. M. Tahsien, H. Karimipour, and P. Spachos. Machine learning based solutions for security of internet of things (iot): A survey. *Journal of Network and Computer Applications*, 161:102630, 2020.

[20] T. Team. 51 most used machine learning tools by experts, Jun 2021.

[21] A. Thakkar and R. Lohiya. A review on machine learning and deep learning perspectives of ids for iot: recent updates, security issues, and challenges. *Archives of Computational Methods in Engineering*, 28(4):3211–3243, 2021.

[22] O. Tsymbal. Iot trends to drive innovation for business in 2022, Apr 2022.

[23] A. Verma and V. Ranga. Machine learning based intrusion detection systems for iot applications. *Wireless Personal Communications*, 111(4):2287–2310, 2020.

[24] N. Waheed, X. He, M. Ikram, M. Usman, S. S. Hashmi, and M. Usman. Security and privacy in iot using machine learning and blockchain: Threats and countermeasures. *ACM Computing Surveys (CSUR)*, 53(6):1–37, 2020.

[25] M. Zolanvari, M. A. Teixeira, and R. Jain. Effect of imbalanced datasets on security of industrial iot using machine learning. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 112–117. IEEE, 2018.

[22] [18] [11] [20] [1]