A Machine Learning Framework for Cybersecurity Operations

Vivek Darak
Department of Information Technology
Pune Institute of Computer Technology, Pune

Mohak Gadge
Department of Information Technology
Pune Institute of Computer Technology, Pune

Shreyash Dhangare
Department of Information Technology
Pune Institute of Computer Technology, Pune

Naman Buradkar
Prof., Department of Information Technology
Pune Institute of Computer Technology, Pune

Abstract:- Compared to the last few decades and past developments in computer and communication technologies along with the internet have provided advanced changes in all of our lives. However, it also opened a whole new frontier for us regarding the security of the system. For example, the privacy of personal information, the security of stored data, availability of stored information, etc. Ensuring the cybersecurity of an enterprise is the work of SIEM systems (Software Information and Event Management). At the SIEM level, the system provides the report regarding the malicious user's intrusion attempts as well as any other dangerous activities on the system. Many of these alerts are however false and are not that dangerous to be avoided so that the prior and important issues of the system are faced like intrusion detection and vulnerable ports. Machine Learning can effectively help us in analyzing the system throughout all the safety parameters to detect all the threats on the system and classify them according to the severity of the alert as well as the frequency at which that particular alert is arriving at the system.

Keywords:- *Machine Learning, Cybersecurity, Intrusion Detection, Software Information And Event Management, Risky User Detection.*

I. INTRODUCTION

Outstanding growth and utilization of the issue of the net increase approximately a way to speak and guard the virtual data safely. In ultra-modern global hackers use distinctive kinds of assaults for purchasing the treasured data. Many of the intrusion detection strategies, techniques and algorithms assist to locate the ones numerous assaults. Malicious assaults have come to be extra advanced and the best assignment is to discover unknown and confused malware because the malware authors use distinctive evasion strategies for data concealing to save you detection through an IDS. In addition, there was a boom in safety threats which include zero-day assaults designed to goal net users. Hence, laptop safety has come to be vital as the usage of data generation has come to be a part of our everyday lives. Cybersecurity is vital as it encompasses the entirety that relates to shielding our touchy facts, in my opinion, identifiable data(PII), covered fitness data (PHI), non-public data, highbrow property, facts, and governmental and enterprise data structures from robbery and harm tried through criminals and adversaries. In a previous couple of decades, device studying has been used to enhance intrusion detection, and presently there's a want for an up-to-date, thorough taxonomy and survey of this current work. There is a massive range of associated research the usage of both the KDD-Cup ninety-nine or DARPA dataset to validate the improvement of Intrusion Detection Systems, but there's no clean solution to the query of which facts mining strategies are extra effective. And after that, the time taken for constructing IDS isn't taken into consideration withinside the assessment of a few IDSs strategies, no matter being an important issue for the effectiveness of IDS.

II. SCOPE

1) To design and develop a framework for detection of intrusion, malware on any application.
2) To predict what attacks are possible on the application.
3) To make an application less vulnerable to cyber-attacks. 4) To automate SOC (Security Operations Centre) activities.
5) To reduce false-positive signals generated by SIEM (Software information and event management) systems using modern algorithms.
6) To generate a meaningful representation of reports and would provide insights to the user for making applications more secure.

III. LITERATURE SURVEY

Ansam Khraisat and Joarder Kamruzzaman [1] introduced in their work, Cyber-assaults are getting extra sophisticated and thereby offering growing demanding situations as it should be detecting possible intrusions.In case of failure to save you the intrusions should decay the credibility of protection services, e.g. records confidentiality, integrity, and availability. Number of intrusion detection strategies were proposed withinside the literature to address pc protection threats, which may be extensively categorized into Signature-primarily based totally Intrusion.

Detection Systems (SIDS) and Anomaly-primarily based totally Intrusion Detection Systems (AIDS). The survey paper offers a taxonomy of cutting-edge IDS, a complete evaluation of tremendous latest works, and an outline of the datasets generally used for assessment purposes. It additionally offers evasion strategies utilized by attackers to keep away from detection and discusses destiny studies demanding situations to counter such strategies with a view to making pc systems extra secure.

Divyatmika, Manasa Sreekesh [2] have presented in their work, Intrusion detection systems are systems that could hit upon any sort of malicious attacks, corrupted records or any sort of intrusion that could pose risk to the systems.

In the paper, they gave a unique technique to construct a community-primarily based totally intrusion detection machine the usage of system studying technique. They have proposed a simple two-tier architecture to hit upon intrusions on network. Network performance may be labelled as misuse detection and anomaly detection. As the evaluation relies upon the community behaviour, they've taken into consideration records packets of TCP/IP as our enter records. After processing the records with the aid of using parameter filtering, they constructed an independent version on education to set the usage of hierarchical agglomerative clustering. Further, records receive labelled as everyday visitors' sample or intrusions the usage of K-Nearest Neighbour classification. Misuse detection is carried out using the usage of MLP set of rules. Anomaly detection is carried out the usage of Reinforcement algorithm wherein community marketers' study from the surroundings and make choices accordingly. The true-positive charge of the proposed architecture is 0.99 and the fake wonderful charge is 0.01. Thus, the structure gives an excessive degree of safety with the aid of using presenting excessive TP and occasional false positive rate. And, it additionally studies the usual network styles and learns (to construct independent machines) to split ordinary and new data and possible threats.

Jonathon Schwartz [3] has presented in their work, To try to deal with the need for taking advantage of models, this undertaking investigated the software of Reinforcement Learning (RL) to automated penetration testing. RL is an AI optimization method that has the key gain that it does now no longer require a version of the surroundings on the way to produce an attack policy, and alternatively learns the quality coverage via interaction with the surroundings. In the primary level of this take a look at the designed and constructed a fast, lightweight and open-source network attack simulator that may be used to train and take a look at autonomous agents for penetration testing. They did this through framing penetration testing as a Markov Decision Process (MDP) with the recognized configuration of the network as states, the to be had scans and exploits as moves, the praise decided through the fee of machines at the network and the usage of non-deterministic actions to model the results of scans and exploits against machines.

Kathleen Goeschel [4] has presented in their work, Intrusion detection systems monitor network or host packets in an attempt to stumble on malicious activities on a system. Anomaly detection systems have fulfilment in exposing new attacks, typically stated as zero day attacks, but have excessive false-positive rates. False-positive events arise while a pastime is flagged for research but it becomes decided to be benign upon analysis. Computational power and valuable sources are wasted while beside the point statistics are processed, data flagged, analyst alerted, and the irrelevant statistics are eventually disregarded. In an attempt to make IDS greater green the fake fine price needs to be reduced. Their paper proposes a version for reducing FPs using data mining techniques by combining support vector machines, decision trees, and Naïve Bayes.

We construct this machine (figure-1) with the usage of a bottom-up clustering technique as it collects information /patterns/sequences after which successively merges the information into large clusters. In those methods, clusters are constructed up through combining, primarily based totally upon their proximity, current clusters.

IV. METHODOLOGY

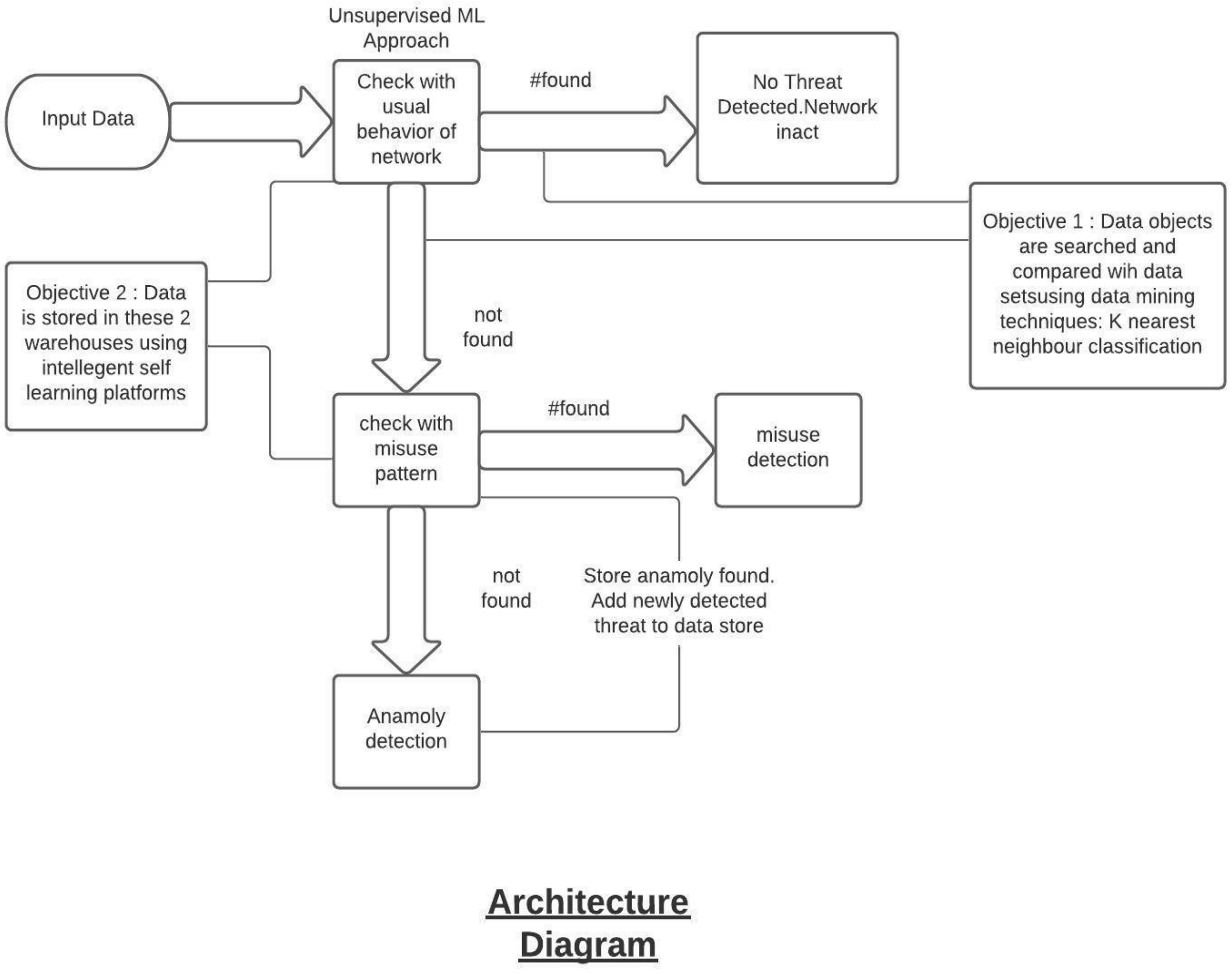
Unsupervised ML Approach
Input Data
Check with usual behavior of network
#found
No Threat Detected.Network inact
Objective 1 : Data objects are searched and compared wih data setsusing data mining techniques: K nearest neighbour classification
Objective 2 : Data is stored in these 2 warehouses using intellegent self learning platforms
not found
check with misuse pattern
#found
misuse detection
not found
Store anamoly found. Add newly detected threat to data store
Anamoly detection
Architecture Diagram

Fig 1

The target knowledge is searched and compared with a group of predefined rules/sequences exploitation KNN rule. This algorithm initially compares the target data with the same old behaviour of the network and a set of trained tagged data containing info concerning malicious data that are harmful to the system. The primary aim here is to establish a device that analyzes the ordinary trends and behaviour of the network and learns to distinguish typical data and irregular threats progressively.

For Host-based intrusion detection, we tend to monitor on host logs. It can capture the intrusions in phrases of the traits or high-satisfactory of diagnosed attacks or device vulnerabilities.

Here we follow the MLP algorithm to construct a misuse detection version. A multilayer perceptron is a feedforward artificial neural network version that inspects the known facts onto a hard and fast of accurate and correct outputs.

An MLP includes a couple of layers of nodes in a directed graph, with every of the layer that is absolutely and truly linked to the subsequent node. MLP makes use of a supervised learning approach that is known as backpropagation for training the network.

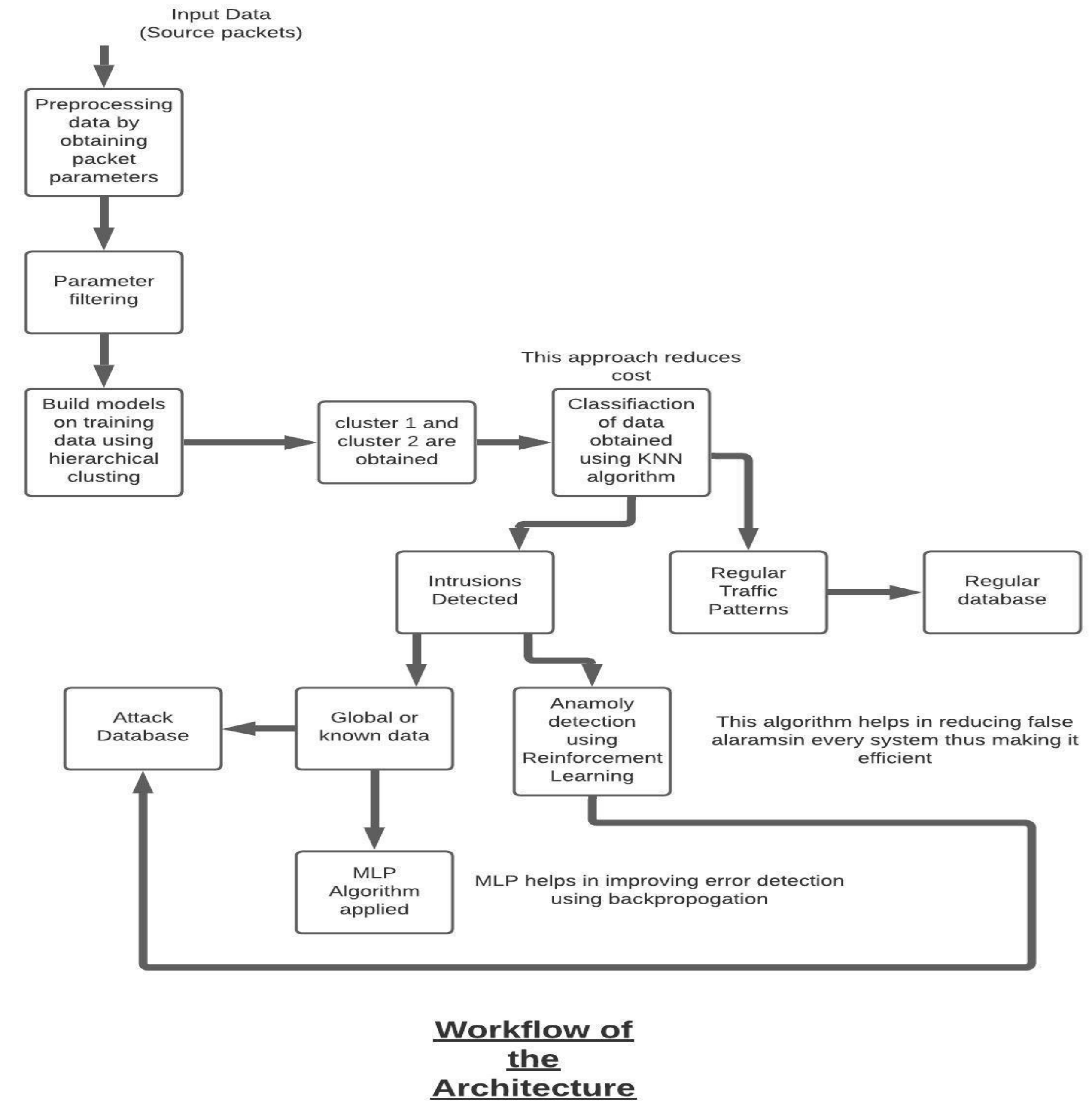
Input Data
(Source packets)
Preprocessing data by obtaining packet parameters
Parameter filtering
Build models on training data using hierarchical clusting
cluster 1 and cluster 2 are obtained
This approach reduces cost
Classifiaction of data obtained using KNN algorithm
Intrusions Detected
Regular Traffic Patterns
Regular database
Attack Database
Global or known data
Anamoly detection using Reinforcement Learning
This algorithm helps in reducing false alaramsin every system thus making it efficient
MLP Algorithm applied
MLP helps in improving error detection using backpropogation
Workflow of the Architecture

Fig 2

Anomaly detection has an excessive stage of fake alarm. So as to face that problem, we observe reinforcement learning where the network is skilled to make choices and predict if any risk exists. This machine (figure-2) makes use of a reinforcement sign dispatched with the aid of using the surroundings to the fusion middle for adjusting the weights defining the selection capacity of every agent and the weights representing trusts in-person choices of every agent. This algorithm decreases the fake alarm rate and the machine no longer wants to waste its sources to address the fake threat.

The surroundings component models network pen-testing problems as an MDP and as such is described with the aid of using the tuple {S, A, R, T}. States are described because of the current knowledge and position of the attacker in the network. Actions are available scans and exploits that the attacker can perform for each machine in the network. The reward feature is simply the value of any machines exploited minus the value of actions performed. The transition function controls the result of any given movement and takes into consideration movement type, connectivity, firewalls and probabilistic nature of exploits.

The action space, A, is the set of available actions in the NAS and consists of a single scan action and can take advantage of every provider and every machine on the network. The reward function is used to define the goals of the autonomous agent and what is trying to be optimized by the agent. The transition function, T, determines how the environment evolves over time as actions are performed.

V. CONCLUSION

We present a user-centric machine learning system which leverages big data of various security logs, alert information, and analyst insights to the identification of malware which may be present in the developed software. Here we describe briefly how to generate labels from SOC investigation notes, to correlate IP, host, and users to generate user-centric features, to select machine learning algorithms and evaluate performances, as well as how to use a machine learning system in the SOC production environment.

REFERENCES

[1]. Khraisat, A., Gondal, I., Vamplew, P. et al. Survey of intrusion detection systems: techniques, datasets and challenges. https://doi.org/10.1186/s42400-019-0038-7.

[2]. Divyatmika and M. Sreekesh, "A two-tier network based intrusion detection system architecture using machine learning approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 42-47, doi: 10.1109/ICEEOT.2016.7755404.

[3]. Autonomous Penetration Testing using Reinforcement Learning by Jonathon Schwartz, Hanna Kurniawati, arXiv:1905.05965

[4]. K. Goeschel, "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis," SoutheastCon 2016, Norfolk, VA, 2016, pp. 1-6, doi: 10.1109/SECON.2016.7506774.