Machine Learning for Cyber Defence

[1]Vaibhav Bhapkar, [2]Prashant Singh, [3]Manish Gaikwad, [4]Arvind Bhat, [5]Urmila Kalshetti
[1]Student, [2]Student, [3]Student, [4]Student, [5]Professor
BE Computer Engineering,
PVG COET (Savitribai Phule Pune University)
Pune, India

Abstract: Machine learning can be used to identify advanced targeting and threats such as organization profiling, infrastructure vulnerabilities and potential interdependent vulnerabilities and exploits. Machine learning can significantly change the cyber security landscape. Malware by itself can represent as many as 3 million new samples an hour. Traditional malware detection and malware analysis is unable to pace with new attacks and variants. New at tacks and sophisticated malware have been able to bypass network and end-point detection to deliver cyber-attacks at alarming rates. New techniques like machine learning must be leveraged to address the growing malware problem. This paper delivers a use of reputation based system for network addresses and machine learning technique to detect and highlight advanced malware for cyber defense analysts.

***IndexTerms* - Dynamic Analysis, Machine Learning Algorithms, Malware Detection, Static Analysis, Support Vector Machine and User Interface.**

I. INTRODUCTION

As computers becomes more ubiquitous the concern for security of these computer implemented systems increases. In this era with such technological advancements one cannot try to not think about the threats and various security aspects of the involved technology.

Also, the advancements in computer networking have not made it any easier for engineers to tackle the security threats. Most common of all the security threats a computer system can face is of Malwares. Malware is a generalized term for all the malicious software which is specifically designed to disrupt, damage, or gain authorized access to a computer system. A common user may or may not be aware of this term but he/she may have come across the following terms associated with it;
Few of the forms of malware are:

1. Adware: The least dangerous and most lucrative Malware. Adware displays ads on your computer.
2. Spyware: Spyware is software that spies on you, tracking your internet activities in order to send advertising (Adware) back to your system.
3. Virus: A virus is a contagious program or code that attaches itself to another piece of software, and then reproduces itself when that software is run. Most often this is spread by sharing software or files between computers.
4. Worm: A program that replicates itself and destroys data and files on the computer. Worms work to "eat" the system operating files and data files until the drive is empty.
5. Trojan: The most dangerous Malware. Trojans are written with the purpose of discovering your financial information, taking over your computer's system resources, and in larger systems creating a "denial-of-service attack" Denial-of-service attack: an attempt to make a machine or network resource unavailable to those attempting to reach it. Example: AOL, Yahoo or your business network becoming unavailable.
6. Rootkit: This one is likened to the burglar hiding in the attic, waiting to take from you while you are not home. It is the hardest of all Malware to detect and therefore to remove; many experts recommend completely wiping your hard drive and reinstalling everything from scratch. It is designed to permit the other information gathering Malware in to get the identity information from your computer without you realizing anything is going on.
7. Backdoors: Backdoors are much the same as Trojans or worms, except that they open a "backdoor" onto a computer, providing a network connection for hackers or other Malware to enter or for viruses or SPAM to be sent.
8. Keyloggers: Records everything you type on your PC in order to glean your login names, passwords, and other sensitive information, and send it on to the source of the keylogging program. Many times keyloggers are used by corporations and parents to acquire computer usage information.
9. Rogue security software: This one deceives or misleads users. It pretends to be a good program to remove Malware infections, but all the while it is the Malware. Often it will turn off the real Anti-Virus software.
10. Ransomware: If you see this screen that warns you that you have been locked out of your computer until you pay for your cybercrimes. Your system is severely infected with a form of Malware called Ransomware. It is not a real notification from the FBI, but, rather an infection of the system itself. Even if you pay to unlock the system, the system is unlocked, but you are not free of it locking you out again. The request for money, usually in the hundreds of dollars is completely fake.

With Governments and Public Organizations spending millions of dollars to protect their systems from attacks, one would wonder what can be done to prevent these attacks. One solution could be to manually check everything, from files to webpages to programs. But this would take a lot of time and resources which makes this solution impossible to implement. This is where machine learning comes into the picture. Now a days, anyone can learn to use machine learning to develop a software or for research purposes.

One can also wonder what is wrong with traditional signature-based and change-based Malware detection methods. The problem is that traditional techniques are not able to cope with new types of malware attacks. With few changes in code the programmer can change the signature of malwares but what does not change is the behavior of the malicious code.

In order to address this problem, in this research project, we plan to develop a practical and efficient way of detecting malwares using machine learning model trained on the behavioral data of the malware files. We study different approaches to solve the above problem and design a system (i.e. architecture) to fulfill our objective.

II. System Overview

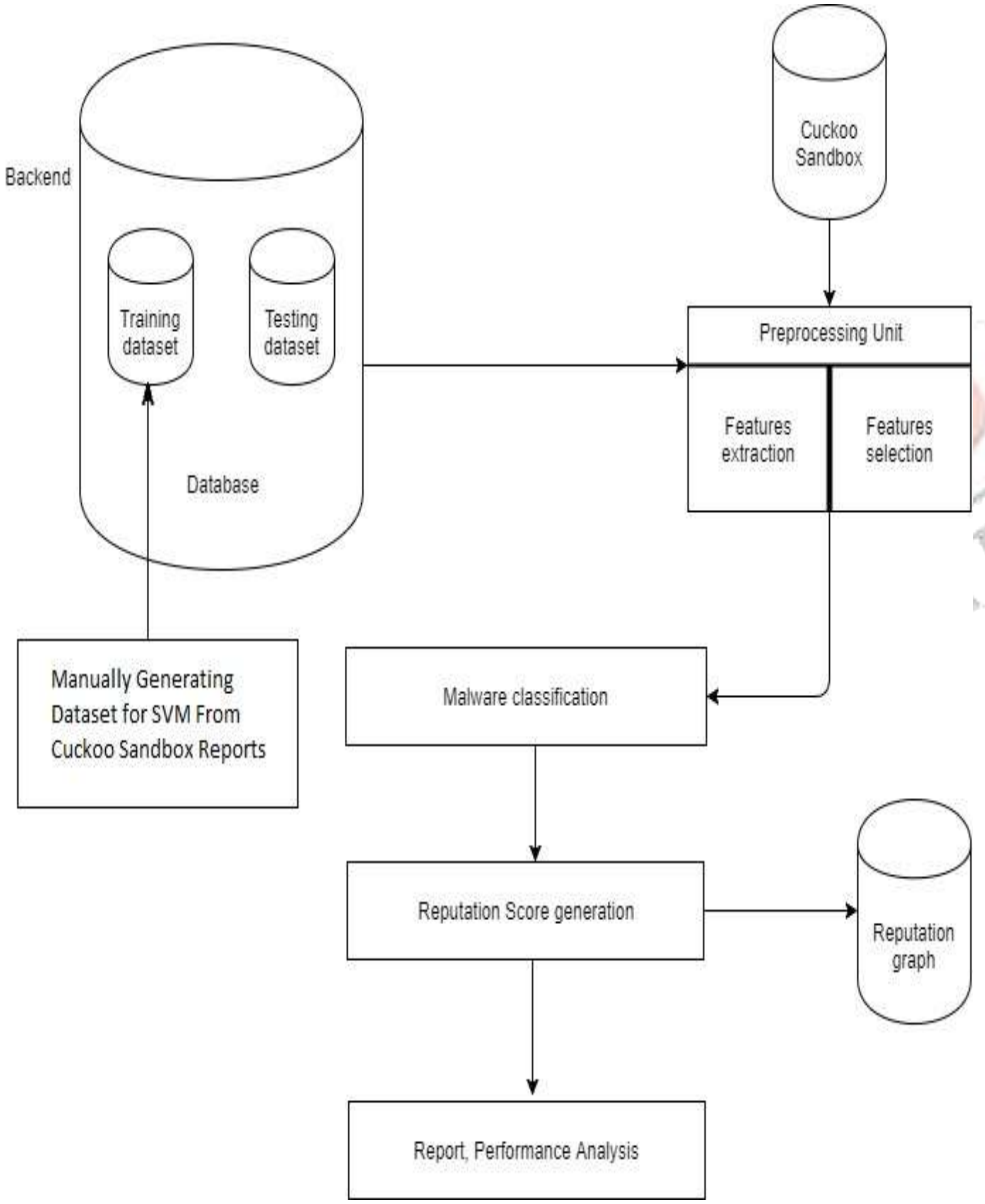

Fig. 1. System architecture

The goal of our project is to develop software for detecting malwares using the concept of reputation and machine learning. We have proposal of system into two main phases:

2.1 Training Phase

The first phase is training of the machine. This requires training SVM machine with a large number of dataset, to predict various malwares with high accuracy and low false positive rates. This generates a training model which is used for further testing. The detailed process used for training of machine learning model is explained using following dataflow training diagram.

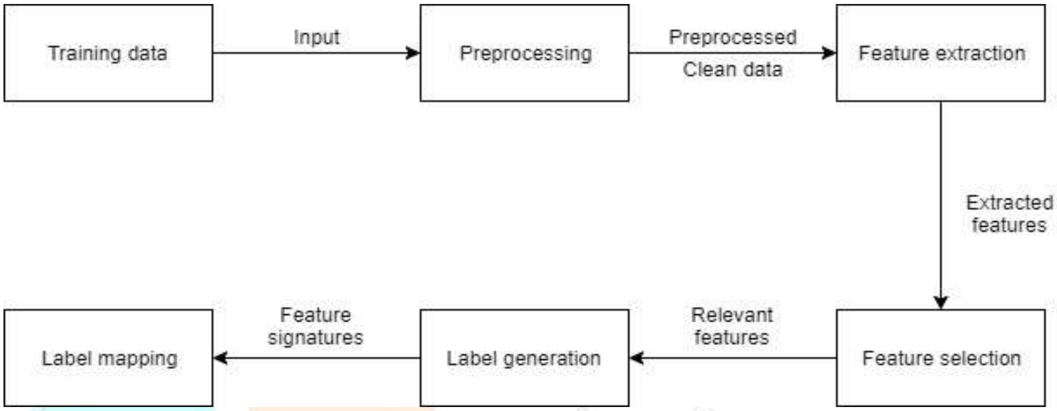

Fig. 2. Dataflow Training

2.2 Testing Phase

The second phase is testing of machine. Which is used to predict the class of malware by performing prediction on previously trained module. The Working of testing phase with reputation is explained using dataflow testing diagram.

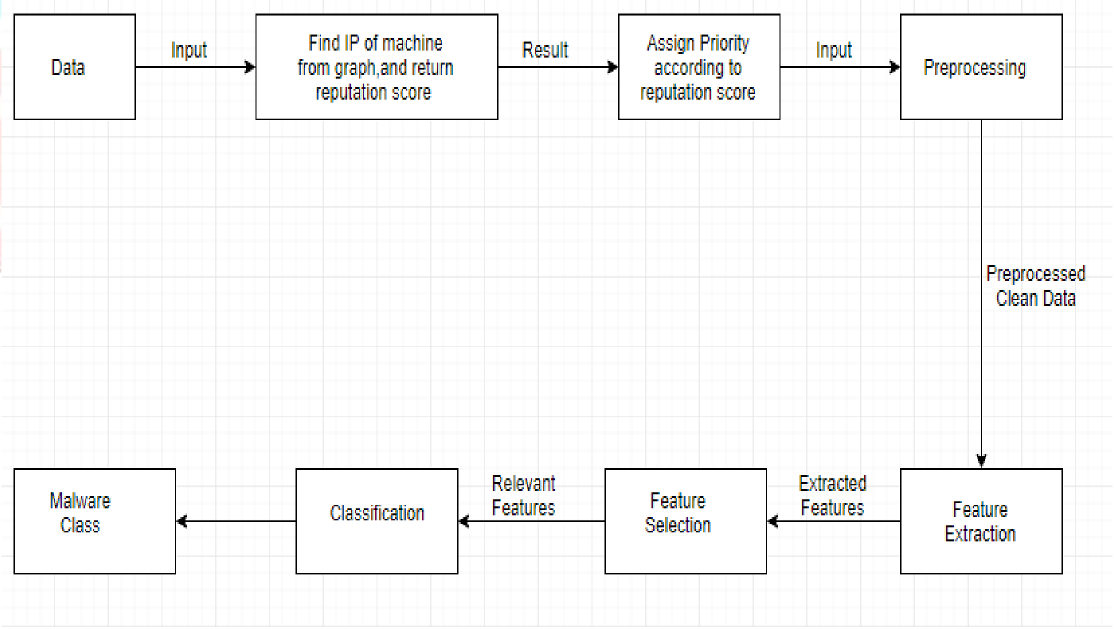

Fig. 3. Dataflow Testing

III. MACHINE LEARNING

The main part while implementing the machine learning for malware detection is preparation of the dataset. Which will consists of mainly two parts,

1. First part is collection of malware samples of various types and analyze that samples into sandbox environment for observing the behavior of that malware.

2. Second part is to extract the features from the report which is generated by analyzing the malware samples and extraction labelling part is important which is used to classify type of malware.

After competing this two steps dataset will be prepared which is used as training set to train classifier of machine learning model the algorithm that we are going to select to classify the model is based on the accuracy of algorithm so we are training our machine learning model on Support vector machine, Random forest and decision tree algorithms to find best result.

IV. Conclusion

Malicious software is one of the major threats in the Internet to day. Many problems of computer security, such as denial-of-service attacks, identity theft, or distribution of spam and phishing contents, are rooted in the proliferation of malware. Several techniques for automated analysis of malware have been developed in the last few years, ranging from static code inspection to dynamic analysis of malware behavior. While static analysis suffers from obfuscation and evasion attacks, dynamic analysis alone requires a considerable amount of manual inspection for crafting detection patterns from the growing amount and diversity of malware variants. In this project, we introduced a framework to overcome this deficiency and enhance the current state-of-threat. Our main contribution is a learning-based framework for the automatic analysis of malware behavior. To apply this framework in practice, it success to collect a large number of malware samples and monitor their behavior using a cuckoo sandbox environment and we are able to apply Machine learning algorithm for the analysis and detection of malwares.

References

[1] Jing, Ranzhe, and Yong Zhang**.**2010**.** A View of Support Vector Machines Algorithm on Classification Problems.

[2] Hung, Pham Van. 2011. An approach to fast malware classification with machine learning technique.

[3] Rieck, Konrad, Philipp Trinius, Carsten Willems, and Thorsten Holz. 2011**.** Automated Analysis of Malware behavior using machine learning.

[4] Alazab, Mamoun, Sitalakshmi Venkatraman, Paul Watters, and Moutaz Alazab. 2011. Zero-day Malware Detection based on Supervised Learning Algorithms of API call Signatures.

[5] Enterprise, Symantec. Internet Security Threat Report 2015.

[6] Reddy, Krishna Sandeep, and Arun Pujari. 2006. N-gram analysis for computer virus detection.