PREVENTION OF CYBER-ATTACKS USING MACHINE LEARNING ALGORITHM

Seminar Report submitted in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

By

Mr. SHUBHAM SANTOSH UPADHYAY (17UECN0074)
Mr. SUDHEER KUMAR GUPTA (17UECS0730)
Mr. VASUJIT BHATTACHARJEE (17UECS0797)

Under the guidance of

Dr. V. Srinivasa Rao M.Tech., Ph.D.,
PROFESSOR & HEAD

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SCHOOL OF COMPUTING

VEL TECH RANGARAJAN Dr. SAGUNTHALA R & D INSTITUTE OF SCIENCE AND TECHNOLOGY, CHENNAI 600 062, TAMILNADU, INDIA
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

October, 2019

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

This is to certify that the seminar entitled "**PREVENTION OF CYBER-ATTACKS USING MACHINE LEARNING ALGORITHM**" submitted by SHUBHAM SANTOSH UPADHYAY (17UECN0074), SUDHEER KUMAR GUPTA (17UECS0730) and VASUJIT BHATTACHARJEE (17UECS0797) in partial fulfillment for the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering is an authentic work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the project report has not been submitted to any other University/Institute for the award of any Degree or Diploma.

Signature of Supervisor

Dr. V. Srinivasa Rao M.Tech.,Ph.D.,
Professor,
Head of Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R & D Institute of Science and Technology,
Avadi, Chennai-600062

Signature of Seminar Handling Faculty

Dr. Carmel Mary Belinda M.E.,Ph.D.,
Associate Professor,
Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology,
Avadi, Chennai-600062.

Submitted for the partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering from Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology (Deemed to be University, u/s 3 of UGC Act,1956).

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Shubham Santosh Upadhyay
17UECN0074)

(Sudheer Kumar Gupta
17UECS0730)

(Vasujit Bhattacharjee
17UECS0797)

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

APPROVAL SHEET

The Seminar report entitled **PREVENTION OF CYBER ATTACKS USING MACHINE LEARNING ALGORITHM** by **SHUBHAM SANTOSH UPADHYAY, SUDHEER KUMAR GUPTA** and **VASUJIT BHATTACHARJEE** is approved for the degree of B.Tech Computer Science and Engineering.

Signature of Supervisor

Dr. V. Srinivasa Rao M.Tech.,Ph.D.,
Professor,
Head Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R & D Institute of Science and Technology,
Avadi, Chennai-600062

Signature of HEAD & DEAN, SOC

Dr. V. Srinivasa Rao M.Tech.,Ph.D.,
Professor,
Head Department of CSE,
Vel Tech Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology,
Avadi, Chennai-600062.

Date :_____

Place:_____

ACKNOWLEDGEMENT

We express our deepest gratitude to our respected **Founder Chancellor and President Col. Prof. Dr. R. RANGARAJAN B.E. (EEE), B.E. (MECH), M.S (AUTO). DSc., Foundress President Dr. R. SAGUNTHALA RANGARAJAN M.B.B.S.**, Chairperson Managing Trustee and Vice President.

We are very much grateful to our beloved **Vice Chancellor Prof. V.S.S. KUMAR, Ph.D.,** for providing us with an environment to complete our project successfully.

We are obligated to our beloved **Registrar Mrs. N.S PREMA.,** for providing immense support in all our endeavors.

We are thankful to our esteemed **Director Academics Dr. ANNE KOTESWARA RAO, Ph.D.,** for providing a wonderful environment to complete our project successfully.

We record indebtedness to our **Head of the Department/Dean Dr. V. SRINIVASA RAO., M.TECH, Ph.D.,** for immense care and encouragement towards us throughout the course of this project.

We also take this opportunity to express a deep sense of gratitude to Our **Internal Supervisor Dr. V. SRINIVASA RAO., M.TECH, Ph.D.,** for his cordial support and guidance, he helped us in completing this project through various stages.

A special thanks to our **Seminar Coordinator Mrs. S. HANNAH., M.E, Ms. D. FEMI, M.E., Mrs. T. ANJALI, M.E.,** for their valuable guidance and support throughout the course of the project.

We thank our **Seminar handling Faculty Dr. Carmel Mary Belinda M.E., Ph.D.,** for the valuable information shared on proceeding with our seminar.

We thank our department faculty, supporting staffs, parents, and friends for their help and guidance to complete this seminar.

SHUBHAM SANTOSH UPADHYAY (VTU10425) (17UECN0074)
SUDHEER KUMAR GUPTA (VTU10449 (17UECS0730)
VASUJIT BHATTACHARJEE (VTU10453) (17UECS0797)

TABLE OF CONTENTS

ABSTRACT

The ML field, which can be briefly defined as enabling computers to make successful predictions using past experiences, has exhibited an impressive development recently with the help of the rapid increase in the storage capacity and processing power of computers. Nowadays, a large amount of data is available everywhere. Therefore, it is very important to analyze this data in order to extract some useful information and to develop an algo. based on this analysis. This can be achieved through data mining and ML. ML is used to design algo based on data trends and historical relationships between data. diverse ML methods will be successfully deployed to address such wide-ranging problems in computer security This will discusses and highlight different applications of ML in cyber security.

CHAPTER 1

INTRODUCTION

Cybersecurity is a major concern for a large number of organizations, institutions, corporations and individuals across the globe. The totality of the technologies and processes for the monitoring and prevention of unauthorized access, alteration, misuse, and denial of service to computer networks and resources constitute cybersecurity. This also includes the tendency to authorize access to classified contents, and critical infrastructure, which are network-accessible. Most networks are widely connected to each other through the Internet, and provide a means for sharing data, information, intelligence, software, and hardware. Though the sharing of valuable resources for enhanced operational efficiency has characterized the computer networking paradigm, it has also created a seamless source of easy propagation of malware, and through this, the escalation of cyber-attacks has become prevalent in the cyberspace. this expansion in the threat landscape is a factor of the growing power of cyber force, which is gradually creeping into the control of all domestic, business and industrial functions. As a consequence of the effect of cyber force, the dangers of cyber-attacks come with the ability to modify the parameters of a system using neural network or database in order to generate a kinetic effect for escalating attacks including the tendency to destroy classified contents. Protecting against cyber-attacks requires both proactive and reactive approaches.

These approaches, which can also be described as active and passive are relevant in the context of use – basically direct defensive actions or mitigation techniques against cyber threats. The relevance of cyber defense strategies is embedded in the capacity to truncate active and passive threats, which have become a norm in the cyber domain. It is therefore pertinent to understand the research gaps in the current cybersecurity approaches. To this effect, this paper will highlight the numerous techniques available in the public domain including the strengths and weaknesses of each. Similarly, most attack prevention approaches are achieved through traffic analysis to identify and drop (or block) a malicious activity.

CHAPTER 2

LITERATURE SURVEY

2.1 DETECTION OF DRIVE-BY DOWNLOAD ATTACKS USING MACHINE LEARNING APPROACH

Drive-by download refers to attacks that automatically download malwares to user's computer without his knowledge or consent. This type of attack is accomplished by exploiting web browsers and plugins vulnerabilities. The ultimate goal of drive-by download attack is to take control of the client's system through exploiting the vulnerabilities of web browsers or its extensions forcing it to perform undesirable operations.

SOLUTION

This section provides a description of the proposed approach, dataset collection process, classifiers used for the analysis and the development of the final GUI program. Figure 2.1 shows the flow diagram with the proposed system various stages.The first step is the collection of the URLs dataset that are classified to be benign or malicious according to (Alexa, 2016). Then those URLs are validated to make sure they still exist and will not return default server error page. Next, MATLAB is used to obtain the HTML source code of these URLs. Then the features extractor code parses the returned HTML code to extract the features and generate a spreadsheet file that will be used as an input to the training system. For validation the dataset was divided into 5 folds. We examine the dataset with 23 different classifiers, the results of the testing are sorted using accuracy and the best 5 are chosen. Finally, a GUI program was developed to let the user inspect URLs based on the top 5 classifiers.

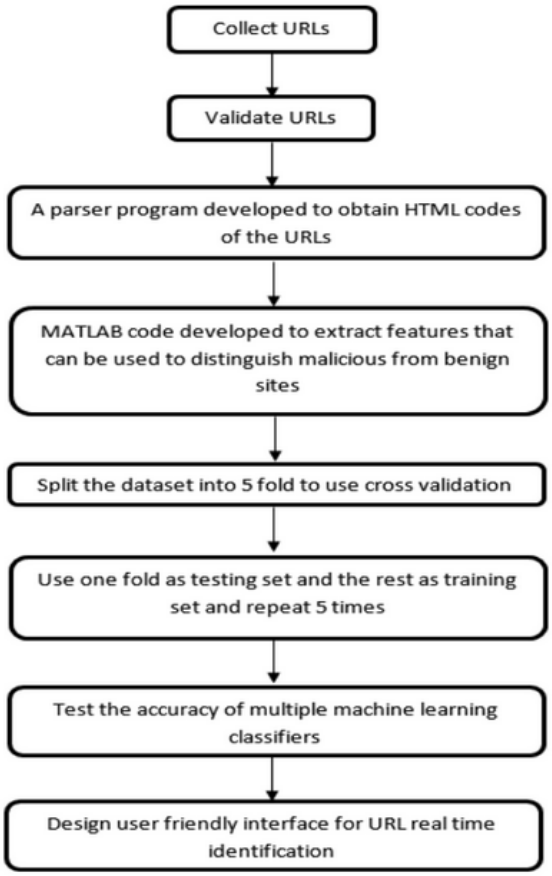


Fig 2.1: Proposed system flow diagram

2.2 MALWARE DETECTION USING MACHINE LEARNING

A versatile framework in which one can employ different ML algo to successfully distinguish between malware files and clean files, while aiming to minimise the number of false positives. In this paper we present the ideas behind our framework by working firstly with cascade one-sided perceptrons and secondly with cascade kernelized one-sided perceptrons. After having been successfully tested on medium-size datasets of malware and clean files, the ideas behind this framework were submitted to a scaling-up process that enable us to work with very large datasets of malware and clean files.

ALGORITHM USED

Algorithm 1 The Perceptron Training Subroutine

```
Sub Train (R, LR_Malware, LR_Clean) :
γ_i = 0
i = 1, . . . , n
for all record in R do
  if Fitness(record) ≠ record.label then
    for all γ_i do
      if record.F_i ≠ 0 then
        if record.label = 1 then
          γ_i = γ_i + LR_Malware
        else
          γ_i = γ_i + LR_Clean
        end if
      end if
    end for
  end if
end for
for all w_i do
  w_i = w_i + γ_i
end for
End sub
```

F = (fa1, fa2, . . . , fan) is an array representing the feature values associated to a file, where fai are file features.

Ri = (Fi, labeli) is a record, where Fi is an array of file feature as above, and labeli is a boolean tag. The value of labeli identifies the file characterised by the array of feature values Fi as being either a malware file or a clean file.

R = (R1, R2, . . . Rm) is the set of records associated to the training files that we use. Algorithm 1 is the the standard perceptron algorithm. Instead of working with floats, it uses a large integer representation for the weights wi , i = 1 . . . n, where n is the total number of attributes/features.

Algorithm 2 One-Sided Perceptron

$NumberOfIterations \leftarrow 0$
$MaxIterations \leftarrow 100$
repeat
  Train (R, 1, -1)
  while FP(R) > 0 **do**
    Train (R, 0, -1)
  end while
  $NumberOfIterations \leftarrow NumberOfIterations + 1$
until (**TP**(R) $= NumberOfMalwareFiles$) or
    ($NumberOfIterations = MaxIterations$)

There are two steps inside the repeat loop of Algorithm 2. The first step, Train(R, 1, -1), performs usual training on the labeled data, obtaining a linear separator (see the perceptron algorithm). The second step, while FP(R) > 0 do Train(R, 0, -1), tries to further move the linear separator, until no clean file is eventually misclassified.

d F = (fa1, fa2, ..., fan). We map F to F ′ so that F ′ = (f ′ a1 , f′ a2 , ..., f′ am), where m = n(n + 1)/2 and f ′ ak = fai&faj , i = [k/n] + 1, j = k%n + 1, k = 1 . . . m, where & denotes the logical and operator.

Algorithm 3 Simple Feature Generation

$pos \leftarrow 0$
for i = 1 to n **do**
  for j = 1 to n **do**
    $f'_{apos} \leftarrow f_{a_i} \& f_{a_j}$
    $pos \leftarrow pos + 1$
  end for
end for

The number of resulted features in F ′ will be n(n + 1)/2, where n is the number of features in F. The computational time increases heavily (e.g. for 308 features in F, we will have 47586 features in F ′ . However, the detection rate (i.e sensitivity) at cross-validation increases with about 10%, as it will be shown later in the results section.

Finally, we used the same one-sided perceptron , but in the dual form and with the training entry mapped into a larger feature space via a kernel function K.

Algorithm 4 Kernelized One-Sided Perceptron

for i = 1 to n **do**
  $\Delta_i \leftarrow 0$
  $\alpha_i \leftarrow 0$
end for
for i = 1 to n **do**
  if $(label_i \times \sum_{j=1}^{n}(\alpha_j \times K(i,j))) \leq 0$ **then**
    $\Delta_i \leftarrow \Delta_i + label_i$
  end if
end for
for i = 1 to n **do**
  $\alpha_i \leftarrow \alpha_i + \Delta_i$
  $\Delta_i \leftarrow 0$
end for

The algorithms 1, 2 and 4 presented above will be used in the sequel as bricks in cascade (or: multi-stage) classification algorithms. Given a set of binary classification algorithms {A1, A2, . . . , Ak}, a cascade over them is an aggregated classification algorithm that classifies a given test instance x.

Algorithm 5 Cascade Classification

if $A_{i_1}(x)$ or $A_{i_2}(x)$ or $\ldots A_{i_k}(x)$ **then**
  return 1
else
  return -1
end if

We performed cross-validation tests by running the three versions of the cascade one-sided perceptron presented on the training dataset.

For the kernelized one-sided perceptron, the following kernel functions were used:

• Polynomial Kernel Function:

K(u, v) = (1+ < u, v >) d , where < u, v > denotes the dot product of the u and v vectors;

• Radial-Base Kernel Function:

K(u, v) = exp(−|u−v|22×σ2).

2.3 INTRUSION DETECTION SYSTEM USING ML

Intrusion detection systems have been highly researched upon but the most changes occur in the data set collected which contains many samples of intrusion techniques such as brute force, denial of service or even an infiltration from within a network.

As network behaviors and patterns change and intrusions evolve, it has very much become necessary to move away from static and one-time datasets toward more dynamically generated datasets which not only reflect the traffic compositions and intrusions of that time, but are also modifiable, extensible, and reproducible.

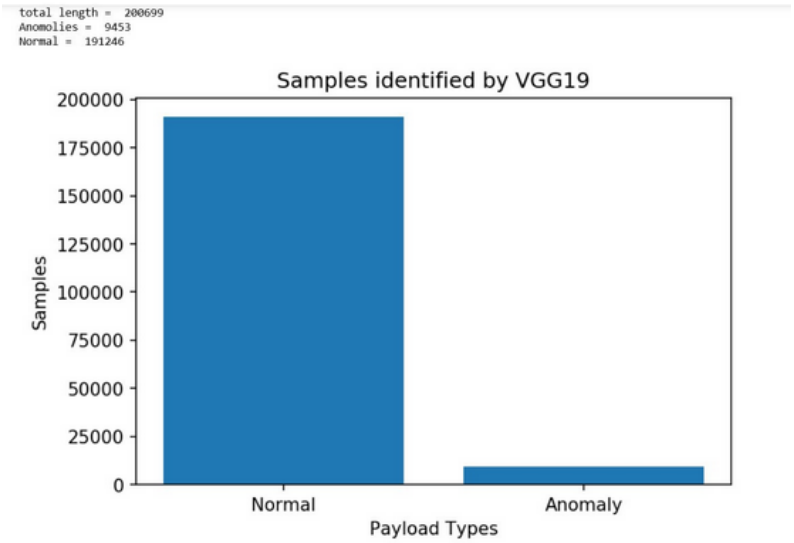


Fig 2.2: Image visualizing the anomaly data from the normal using matplotlib library

2.4 The Model

We tried various experiments with different models but the one we found with successful accuracy was using transfer learning technique on the VGG-19 Keras pre-trained model.

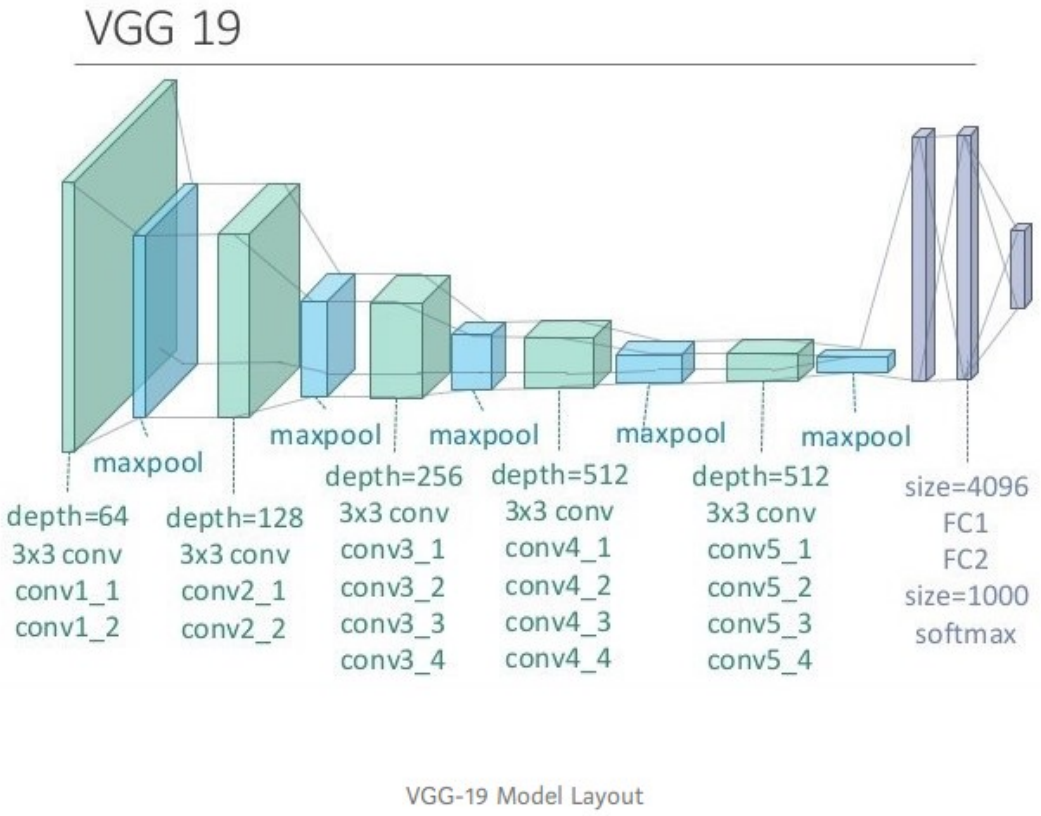
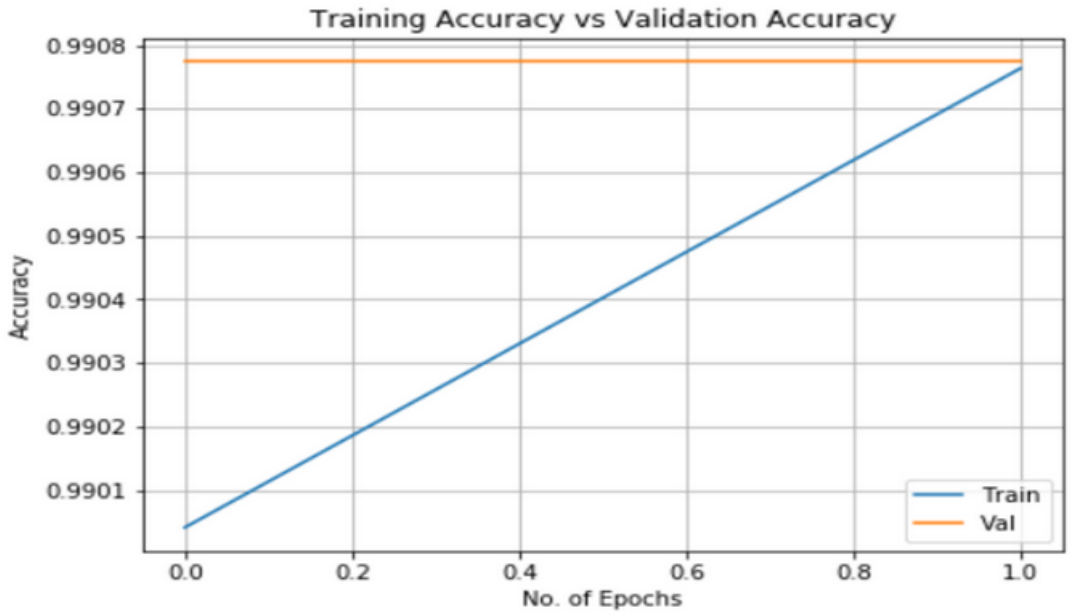


Fig 2.3: VCG-19 Model Layout

Training Phase: The phase where a profile of normal payloads is created. Normally, any anomaly data is not needed for training as the IDS will instantly discard any data with the smallest of deviation from among the data in the profile during the next phase.

1. **Testing Phase:** The phase were incoming payloads are compared with the data stored in the profile.



Accuracy graph of the model in 2 epochs

	precision	recall	f1-score	support
normal	1.00	0.99	1.00	191266
anomaly	0.85	0.98	0.91	9433
avg / total	0.99	0.99	0.99	200699

Fig 2.4: Scientific model evaluation of our model using Scikit-learn library

CHPTER 3

METHODOLOGIES

In today's technology driven world, cyber-attacks are more active than before and putting the average home user at risk. There are many different ways for to protect information from cyber-attacks. Maintaining computer privacy takes a multi-pronged approach. It can be challenging for home users since they don't have much knowledge about cyber-attacks. The quick reference will give more in-depth tips on how to protect your information and significantly reduce the chance of such cyber-attacks.

3.1 DRIVE-BY DOWNLOAD ATTACKS

When a computer becomes infected with malicious software simply by visiting a website, it's known as a *drive-by download*. The industry calls this type of attack a "drive-by" download because the user doesn't have to stop or click anywhere on the malicious page. Simply viewing the page is enough to cause the infection, which happens in the background and without the user's knowledge or consent.

In a drive-by download attack, criminals compromise a website, often a legitimate one, by embedding or injecting malicious objects inside the web pages. The infections are invisible to the user, and range from malicious JavaScript code to iFrames, links, redirects, malvertisements, cross-site scripting, and other malicious elements.

When a user visits an infected web page, the user's browser automatically loads the malicious code, which immediately scans the victim's computer for security vulnerabilities in the operating system and other applications.

3.2 EXISTING SOLUTION THROUGH MACHINE LEARNING:

A data set of 5435 webpages is used for training and testing of 23 different ML classifiers and based on the detection accuracy we selected the top five to build our detection model. The approach is expected to serve as a base for implementing and developing anti drive-by download browser extensions.

3.3 MALWARE ATTACKS

Malware attacks are becoming very common. Every system is threatened of malicious attacks because of different viruses and malwares. However, one can avoid any damage to a system by knowing how malware attacks a system and the way it spreads in it. The term malware is a broad term encompassing Trojan horse virus, worms and other system viruses. Whenever you establish an internet connection for reading your mails or sharing files over the web, your system is exposed to malware attacks.

3.4 EXISTING SOLUTION THROUGH MACHINE LEARNING:

A versatile framework in which one can employ different ML algo to successfully distinguish between malware files and clean files, while aiming to minimise the number of false positives. In this paper we present the ideas behind our framework by working firstly with cascade one-sided perceptrons and secondly with cascade kernelized one-sided perceptrons.

3.5 INTRUSION DETECCTION USING MACHINE LEARNING

As network behaviors and patterns change and intrusions evolve, it has very much become necessary to move away from static and one-time datasets toward more dynamically generated datasets which not only reflect the traffic compositions and intrusions of that time, but are also modifiable, extensible, and reproducible.

Due to the application of ML within the system, anomaly-based detection is rendered the most effective among the intrusion detection systems as they have no need to search for any specific pattern of anomaly, but they rather just treat anything that does not match the profile as “Anomalous”.

3.5.1 NETWORK INTRUSION DETECTION SYSTEM

A network-based intrusion detection system (NIDS) is used to monitor and analyze network traffic to protect a system from network-based threats. A NIDS reads all inbound packets and searches for any suspicious patterns. When threats are discovered, based on its severity, the system can take action such as notifying administrators, or barring the source IP address from accessing the network. Intrusion detection systems (IDSs) are available in different types; the two main types are the host-based intrusion system (HBIS) and network-based intrusion system (NBIS). Additionally, there are IDSs that also detect movements by searching for particular signatures of well-known threats. An IDS compliments, or is part of, a larger security system that also contains firewalls, anti-virus software, etc. A NIDS tries to detect malicious activity such as denial-of-service attacks, port scans and attacks by monitoring the network traffic.

3.5.2 INTRUSION DETECTION SYSTEM (IDS)

Attacks on the computer infrastructures are becoming an increasingly serious problem.An intrusion is defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource. Intrusion detection is therefore required as an additional wall for protecting systems. Intrusion detection is useful not only in detecting successful intrusions, but also provides important information for timely counter measures. Intrusion detection is classified into two types: misuse and anomaly detection. Misuse intrusion detection uses well-defined patterns of the attack that exploit weaknesses in system and application software to identify the intrusions. These patterns are encoded in advance and used to match against the user behavior to detect intrusion. Anomaly intrusion detection uses the normal usage behavior patterns to identify the intrusion. The normal usage patterns are constructed from the statistical measures of the system features. The behavior of the user is observed and any deviation from the constructed normal behavior is detected as intrusion. Dorothy Denning proposed the concept of intrusion detection as a solution to the problem of providing a sense of security in computer systems. The basic idea is that intrusion behavior involves abnormal usage of

the system. Different techniques and approaches have been used in later developments. Some of the techniques used are statistical approaches, predictive pattern generation, expert systems, keystroke monitoring, state transition analysis, pattern matching, and data mining techniques.

CHAPTER 4

RESULTS AND DISCUSSIONS

As we have introduced ML model will be able to identify unusual traffic on the network, and shut down these connections as the occur. A well-trained model would also be able to identify new samples of malware that can evade human generated signatures, and perhaps quarantine these samples before they can even execute. In addition, a ML model trained on the standard operating procedure of a given endpoint may be able to identify when the endpoint itself is engaging in odd behaviour, perhaps at the request of a malicious insider attempting to steal or destroy sensitive information.The ability to collect and handle big data, along with increased ability to perform previously impossible calculations, are significant achievements in the field of ML and hence more cyber attacks can be detected and prevented.

What we are trying to do is, combine most of the already existing ML algo into one and try to solve all the problems in a single go.Given these two points, ML techniques are a great fit to improve the security posture of an organization. And in fact, there are probably ML approaches implemented at some level in your organization.

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 CONCLUSION

ML approaches are a better option for effective tool that can be employed in many areas of information security. There exist some robust anti-phishing algo and network intrusion detection systems. ML can be successfully used for developing authentication systems, evaluating the protocol implementation, assessing the security of human interaction proofs, smart meter data profiling, etc. There are many opportunities in information security to apply ML to address various challenges in such complex domain.

5.2 FUTURE ENHANCEMENT

As ML assumes increased importance in business applications, there is a strong possibility of this technology being offered as a Cloud based service known as ML as a Service (MLaaS) Connected AI systems will enable ML algo to “continuously learn,” based on newly emerging information on the internet. ML will help machines to make better sense of context and meaning of data. Fraud detection, malware detection, intrusion etection, scoring risk in a network, and user/machine behavioral analysis are the five highest ML use cases for improving cybersecurity.

Appendix 1

LIST OF FIGURES

Appendix 2

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPLANATIONS
ML	Machine Learning
IDS	Intrusion Detection System
NIDS	Network Intrusion Detection System
NBIS	Network-Based Intrusion System
AGLO	Algorithm
HBIS	Host-Based Intrusion System
UML	Unified Modelling Language

REFERENCES

[1]. E.Konstantinou,“Metamorphic virus: Analysis and detection,”2008,Technical Report RHUL-MA-2018-2, Search Security Award M.Sc. thesis, 93 pages

[2]. H. Weiwei., and Y. Tan. Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. 2017, February 20, Retrieved: November 03, 2018, from https://arxiv.org/abs/1702.05983v1.

[3]. J. Saleem, B. Adebisi, R. Ande, and M. Hammoudehs A state of the art survey Impact of cyber attacks on SME's. In Proceedings of the International Conference on Future Networks and Distributed Systems (p. 52). ACM, 2017, July..

[4]. M. P. Stoecklin. DeepLocker: How AI Can Power a Stealthy New Breed of Malware. 2018, August 13. Retrieved: September 20, 2018, from https://securityintelligence.com/deeplocker-how-ai-can-powera-stealthy-new-breed-of-malware/

[5]. S. Dolev and S. Lodha, “Cyber Security Cryptography and Machine Learning“, In Proceedings of the First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017.