TABLE OF CONTENTS

ABSTRACT

Internet is a most wonderful tool that is created by a human being. All the information that we require is just a few clicks away. Yet, it is a wonderful tool there are a lot of security issues associated with the internet. A firewall is a tool that prevents applications from cyber security attacks. Yet, there are powerful firewalls still, cyber-attacks are happening around the world. Most of the cyber-attacks going on in the world are due to man-made errors.

So, it is important for us to understand there is a need for a powerful yet fast processing firewall that is driven by machine learning algorithms in order to improve security. There is a common pattern in attacking a web application or web server by hackers. The same pattern can be used to train a machine learning model and add it to a web application to attain maximum security.

A logistic regressing machine learning model is more fitted for these types of machine learning applications which can be trained and tested against the Kaggle dataset. Kaggle dataset consists of more than 12 lakhs security data which also has data from previous cyber attacks.

ML model which is being used in this project is built on sklearn framework and web-part is fully built on the nodejs framework which are Java script frameworks most popularly used for the backend process in web development.

I will be using HTML, CSS, and bootstrap for the front end to create a user interface for users, Nodejs for the backend process and flask for the automation process to get real-time security in the web.

CHAPTER 1

INTRODUCTION

1.1 PROBLEM STATEMENT:

To develop a firewall using machine learning that can detect web attacks efficiently and fastly. Also, it needs to classify the type of attack and protect the application from being hacked.

1.2 OBJECTIVE:

The main objective of this project is to show how normal firewalls fail to give maximum security to a web application or a network server, There are a lot of web servers that are being attacked by hackers. Many web servers are prone to data breaches and attacks by hackers when they are still secured by firewalls. So it is important to understand that there is a strong need for a firewall, which can secure websites and networks much more secure.

We are using a firewall that is being defined by a few predefined rules during its programming. All the attacks/requests that did not satisfy any of those conditions will be restricted by the firewall. but, there is a strong need for a firewall that uses come common phrases in hacking to get trained if there are any requests that involve any of these phrases in them then those requests will be blocked.

1.3 NECESSITY:

Though there are many companies that use very strong firewall systems and still get attacked by hackers. Few hackers steal critical and confidential data from servers which include contact details and payment data, which not only impact the organization but also people who registered on that site. So there is a strong need for a better firewall. Here ML comes in to give better security.

ML-based firewall got trained using ml model using a strong dataset which is a collection of a very large dataset that consists of more the 12lakh attack payloads which are being used by many hackers in past 15 years which is collected from web server data and classed according to its severity. This firewall which is powered by the ML model monitors network packets or web requests and if any request is found to be causing any harm to the webserver or application firewall will immediately block that request.

1.4 OUTLINE

Request and attacks data from web servers were collected. Used data science techniques to extract needed information and remove unwanted information from web requests and form a dataset from it. Using the dataset to train the ML model with high speed and better accuracy.ML model trained using dataset is tested using few test matrics and scripts. An API is developed to monitor website traffic which is also useful as a firewall.

1.5 WEB DESIGN OVERVIEW:

The web design process started with a visual concept by designing in Figma software. Then, I used HTML and CSS to build the website. HTML handles the basic structure of the page, while CSS handles the style and appearance.

CHAPTER 2

LITERATURE SURVEY

In the years gone by, research on the topic of firewalls using machine learning and deep learning algorithms to detect Cyber attacks on web servers and their analysis have taken place widely. This is due to the demand in understanding the deeper relationship between payload and hacking type, and also the relationship between the payload involved themselves.

ROBSON V. MENDONÇA proposed a deep convocational neural network model that is fast in its way and can classify payloads more accurately that uses deep CNN for faster detection. This model is published in Intrusion Detection System Based on Fast Hierarchical Deep Convolutional Neural Network journal 2021.

TAHA SELIM USTUN proposed random forest and decision tree ML models which is used as a firewall for IEC 61850 which is also called the internet of things(IoT) which is connecting physical devices with the internet. Detection is done using symmetric and asymmetric faults. This method is published in Artificial Intelligence Based Intrusion Detection System for IEC 61850 Sampled Values Under Symmetric and Asymmetric Faults 2021.

Dilara Gümü¸sba¸s proposed a few AI models like data encoders, CNN to detect attacks on a web server. It also discusses various datasets available for training AI models in cyber-attack detection. This paper also discusses the importance of data encoding during the training of AI models for better accuracy. This is published in the journal A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems in 2020.

Dennis Appelt gave an overview of SQL Injection and how it can be very dangerous in its own way. He also suggested an algorithm to detect SQL injection efficiently. His work is published in the paper Behind an Application Firewall, Are We Safe from SQL Injection Attacks? in 2015.

Beibei Li suggested a few measures to secure physical systems from cyber attacks. Physical systems in industries are important for the industry securing them from cyber attacks is crucial for industries. Beibei Li also suggested a few measures to maintain privacy in industrial physical systems. This work is published in DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems in 2020.

Gustavo Betarte used a multi-class classifier to detect OWASP top 10 vulnerabilities in web applications. He used a multi-class classifier in the web to parse requests and detect attack/ Hacking payload in the request. He also 3 major datasets to test it. His work is published in Web Application Attacks Detection Using Machine Learning Techniques in 2018.

Dennis Appelt in his paper about testing firewall systems gave an approach to test firewalls in a way that is effective. He also proposed an ML model to generate attack payloads like SQLInjection in testing firewall systems. His work is published in the paper A Machine-Learning-Driven Evolutionary Approach for Testing Web Application Firewalls in 2018.

CHAPTER 3

AIM AND SCOPE OF PRESENT INVESTIGATION

3.1 AIM:

The main aim of this project is to make it much more secure than ever. At the same time, it should be fast and effective. Even though almost 100% of servers are being secured by firewalls, there are a lot of web attacks are happening daily and most of the time firewalls are not able to secure networks or web servers from a new type of attack.

Most of the time hackers try to exploit web servers using various web attacking techniques like SQL injection, Cross-site scripting, remote code execution and web server attacks and many more. A firewall that is being used in a web server is trained in a way that it can detect many of these types of attacks. Ml model uses Sklearn logistic regression model which is a simple, yet fast and effective machine learning model for machine learning.

Dataset used in this ML model consists of more than 12lakh data which is collected from web servers and classified according to the superiority of attack. ML model used in this firewall has attained an efficiency of 97.5% which is specifically trained for injection, XSS, RCE and directory transversal based attacks. It also achieved a great speed of predictions of 1,00,000+ requests per second.

3.2 SCOPE OF THE PRESENT INVESTIGATION:

With the reference from a survey by HackTheBox(HTB), A leading hacking practice platform in association with Bugcrowd community, A leading community of ethical hackers throughout the world in the year 2020, came to know that most the biggest drawback which makes breaching servers easy when the same type of attack can be done to many web servers. Till then many individuals and corporate companies started creating AI-based firewalls but no one got satisfied with that because of its slow nature and less the 50% accuracy in detecting a web attack.

Also, they came to know that from the community of developers and hackers, the firewall they are using utilises a lot of computational power and resources like RAM and network from their machine which also is one of the drawbacks of it.

This firewall use as less as 10GB of storage for training and maintenance of the ML model and is used as less than 1GB ram to compute 1lakh+ of predictions. Which is quite small compared to previous firewalls. It can be further trained to detect DDoS attacks and authentication attacks so that web servers are much more secure.

3.3 SYSTEM STUDY

Using traditional firewalls came to know that the biggest drawback which makes breaching servers easy is when the same type of attack can be done to many web servers. Till now so many companies started creating AI-based firewalls but no one got satisfied with that because of its slow nature and less the 50% accuracy in detecting a web attack. The firewall they are using utilises a lot of computational power and resources like RAM and network from their machine which also is one of the drawbacks of it.

CHAPTER 4

EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHMS USED

4.1 INTRODUCTION TO MACHINE LEARNING:

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed.

The name machine learning was coined in 1959 by Arthur Samuel. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible; example applications include email filtering, detection of network intruders, and computer vision.

Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter sub-field focuses more on exploratory data analysis and is known as unsupervised learning.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

4.2 TRAINING THE DATA:

There are basically two widely-used types of training that can be done to create a model:

i. Supervised Learning

ii. Unsupervised Learning

4.2.1 Supervised Learning:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and the desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

4.2.2 Unsupervised Learning:

Unsupervised machine learning is the machine learning task of inferring a function that describes the structure of "unlabeled" data (i.e. data that has not been classified or categorized). Since the examples given to the learning algorithm are unlabeled, there is no straightforward way to evaluate the accuracy of the structure that is produced by the algorithm—one feature that distinguishes unsupervised learning from supervised learning and reinforcement learning.

The type of training used in this model is **SUPERVISED LEARNING.**