Analysis of cyber-attack using network logs

1Avni Saxena, 2Mansi Agarwal, 3Pranshu Garg,4Eshan Sharma, 5Rajni Ranjan Singh

*1,2,3,4,5Department of CSE& IT M.I.T.S. Gwalior, M.P. (India)*

1saxenaavni1998@gmail.com

2mansi.agarwal211@gmail.com

3pranshu.garg98@gmail.com

4eshan.sharma108@gmail.com

5rrsingh@mitsgwalior.in

***Abstract--* A plethora of devices makes up a network and hence there is possibility of various kinds of cyber-attacks. There is a need to detect these attacks and interrupt the system to avoid further damage. This proposed work uses machine learning and neural network approach to design a system which train the machine to classify the attacks and label them. Denial of service attack (DoS), where a number of packets flood the machine and overload it by buffer overflow, Local area network denial etc. is an important concern here. This paper proposes a framework to detect DoS attack using machine learning(ML) where algorithms like k-nearest neighbor, logistic regression, Gaussian naïve bayes, support vector machine etc. are used.**

***Keywords—Network logs, Wireshark, Denial of service attack, Machine learning.***

I. INTRODUCTION

The internet is made of inter connection of many networks which follow standardized protocol, therefor whenever any transaction happens in this network it has to go though many networks and follow protocols to reach a device. Even a small request of a website on internet is transaction of many packets send to and from a computer. These transaction packets are recorded and saved at many stages which are called logs. The concepts used the paper

1. *LOGS*

Logs are the data which the system automatically produce and information about the events are stored on network devices. Log analysis is important because it helps in recognizing suspicious event sequences in network forensics and to determine how our system is behaving and network forensics. There are various types of logs in our system the logs we can use are as followed.

*1)Firewall logs:* Firewall stops and filter all the packets before they reach the computer. Here the logging documents all these packets and save their header information. Firewall logs also record how firewall deals with traffic types.

*2)Access logs:* Record of all the requests for individual files requested from a website including images and other files that get transmitted are logged in Access logs.

*3)Application logs:* Logging of information the of events that occur within a software application are called application logs. This logging is done by every application and therefore vary application to application. These logs include information event, warning and errors too.

*4)Network logs:* This is the file which contains record of the user and process calls, authentication attempts in a system, basically is shows how systems interacts with each other.

1. *WIRESHARK*

Wireshark is open source and works cross-platform network data packet sniffer and analyzer. It captures packets using pcap. It also has Graphical GUI and non-graphical for Linux flavours [3]. We use wireshark to sniff the network logs and analyses them. At a microscopic level, deep packet inspection of packets are done by applying it at various levels (IP, MAC Address, Port No, Time, Time zone, Protocol, Length of Packet, other Info) [3]. These then are used offline too for further investigation.

1. *CYBER ATTACKS*

These are the assaults done by cyber-criminal to harm other network or system. This can be done for various reasons like stealing, manipulating their data or just breaching their network or even to disrupt their system’s working. There are many types of attacks that occur, following is the one addressed in our project:

*1)DOS attack:* Denial-of-service attack (Dos) occurs when the under-attack machine is flooded with superfluous requests which in result overload the system because of which no legitimate requests are fulfilled. This is done to cut the targeted machine from using network resources temporarily or indefinitely disrupting services of machine connected to the internet. Dos can be done by various ways like Internet control message protocol (ICMP) flooding, Local area network denial (LAND) attack, Syn flood and buffer overflow. These all can be detected in this project.

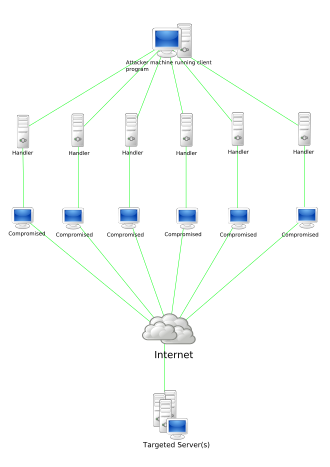


Fig. 1. – DDoS Stacheldraht attack diagram [12]

1. *MACHINE LEARNING*

Machine learning is defined as branch of computer science where different algorithms improve and learn themselves with increasing experience. In Machine learning we create a mathematical model. These models are trained on some training data and then are applied on some additional data to give predictions or decision without actually programmed to do so. In this project too after gathering the logs we perform Machine learning to identify the attacks.

These models are made for the function it has to perform, using different machine learning algorithms. Like the one used in this project are logistic regression, support vector machine, stochastic gradient descent , gaussian naïve Bayes, decision tree, random forest, adaboost and gradient tree boosting.

For the various function in machine learning, libraries are invoked. In earlier times all the coding in machine learning was done manually by using mathematical n statistical formulas but in this time libraries are available with make it easy and efficient to perform machine learning. Following are the few libraries used in this project:

*1)Numpy:* This is a python library which is popularly used for scientific computing, it is used to perform multidimensional array and matrix calculations because of the availability of a huge pool of mathematical functions which are considered as high level.

*2)* *Pandas:* This is another library in python used for data manipulation and analysis. Pandas was developed for data extraction. It provides various tools and high-level data structures and for manipulating time series and numerical tables used for structuring and operations.

*3)* *Matplotlib:* This library is for data visualization and thus is used for displaying 2d graphics, graphs and charts. This Pyplot named module is used to provide features to formatting axes, control line styles etc. and hence it is used for plotting.

*4)* *Seaborn:* This is also a Python library used for data visualization based on matplotlib. As matplotlib is for basic plotting, this provides a variety of visualization patterns to draw informative as well as attractive statistical graphs. These visualization patterns may include high level interface. This also have lesser syntax and many default themes.

*5)* *sklearn:* A free library in python which has several algorithms for performing machine learning. This library contains algorithms for classification, regressions, clustering, dimensionality reduction, model selection and preprocessing. Like support vector machine, random forests, and k-neighbors and etc.

.II. RELATED WORKS

1. Bhavna Dharamkar1, Rajni Ranjan Singh (A Review of Cyber Attack Classification Technique Based on Data Mining and Neural Network Approach )– this work proposed a model which uses neural network as well as data mining in order to classify the different types of cyber-attacks .Data mining is used for the purpose of detection of a cyber-attack using clustering, classification, rule detection whereas feature reduction technique is done by neural network. IDS evaluation is done using DARPA data set and KDD 99 cyber-attack data set.
2. DoS Attack Detection Based on Naive Bayes Classifier V. Hema and C. Emilin Shyni- This proposed work uses Intrusion Detection with the help of Naïve Bayes (IDNB) which is designed to detect intrusion packets for scenarios of attacks where a huge number of packets arrive at a system. This attack is Denial of service attack (DOS attack) which slows the system down by buffer overflow, SYN flood etc. This work also aims at temporarily suspending or interrupting services of the network-connected system.
3. A SURVEY ON WEB APPLICATION ATTACK DETECTION METHODS Kunal Gupta, Rajni Ranjan Singh This work defines various intrusion detection systems (IDS) like signature-based detection, anomaly-based detection, host-based intrusion detection system and network-based system. It also makes use of SNORT IDS and wireshark which is a network data packet analyzer for the purpose of detecting attacks like cross site scripting, DoS attack, DDoS attack.
4. Cyber-Attack Classification using Improved Ensemble Technique based on Support Vector Machine and Neural Network – this work is proposed for the detection of unknown attack using intrusion detection system. It makes use of ensemble methods which use various moderately accurate models for classification and aggregate them into one more accurate model. It proposes a cascaded support vector machine classifier and make use of multiple kernel functions for ensemble classifiers. For feature collection, graph-based technique is also used in this work.

III. WORKING

*A. FRONT END*

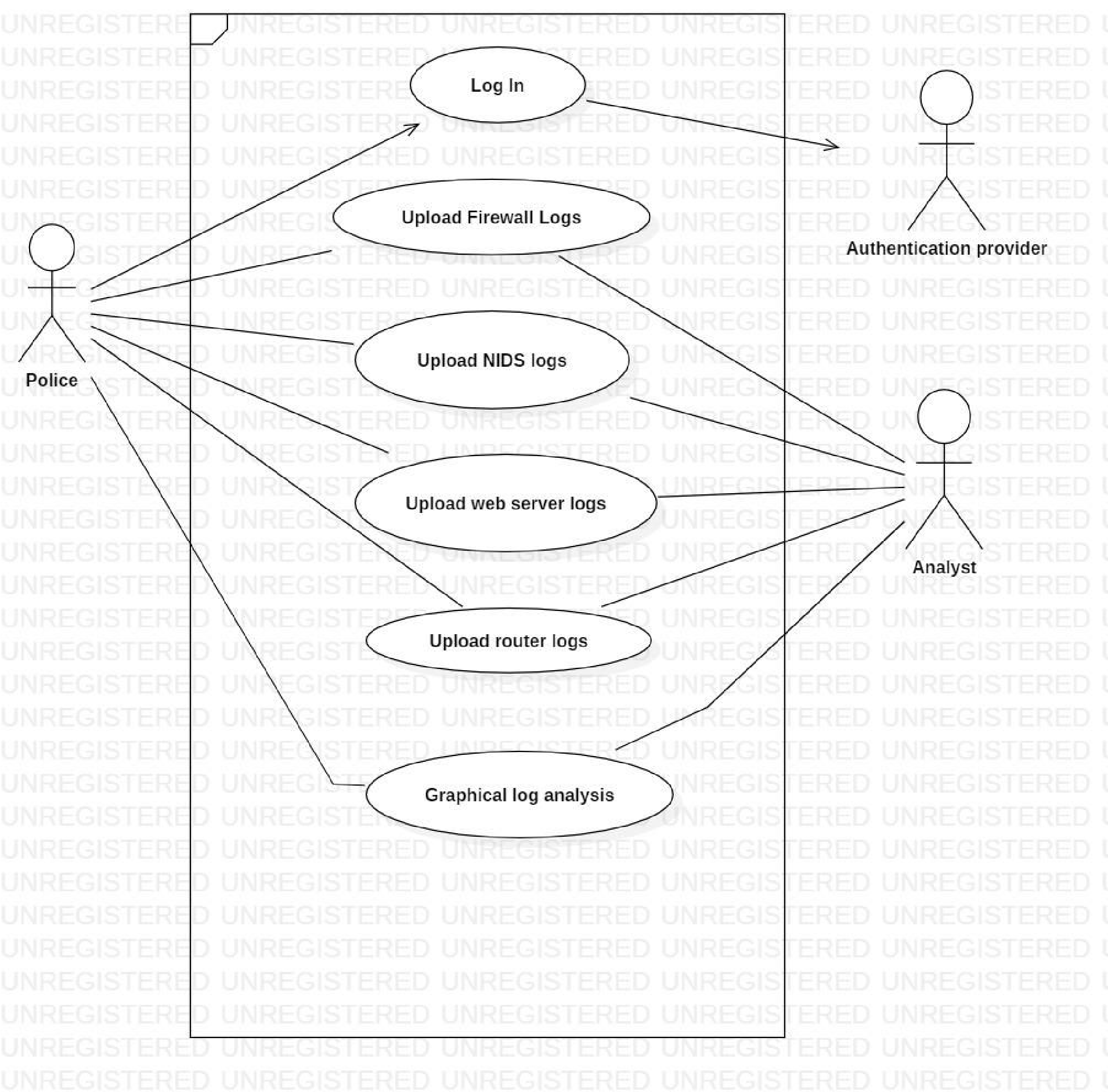


Fig. 2. – Use case Diagram

Lastly, we integrate this machine learning backend with frontend. Front end is designed in python with the help of pyqt5 framework. Pyqt5 is a python package which is used to make an independent application; therefore, our application is also a standalone application which can run on any platform.

In pyqt5 we have used different libraries like QtGui, QtWidgets, QtCore, QFileDialog, QMessageBox.

Our application GUI is divided in 3 main pages:

*1)Log in page:* First we have to login with correct credentials (username and password).

*2)data entry page:* In this page we have to upload the acquired logs (wireshark and firewall logs) in .cvs format.

*3)analysis page:* This page shows the result that the Denial of service attack occurred or not, precision and accuracy.

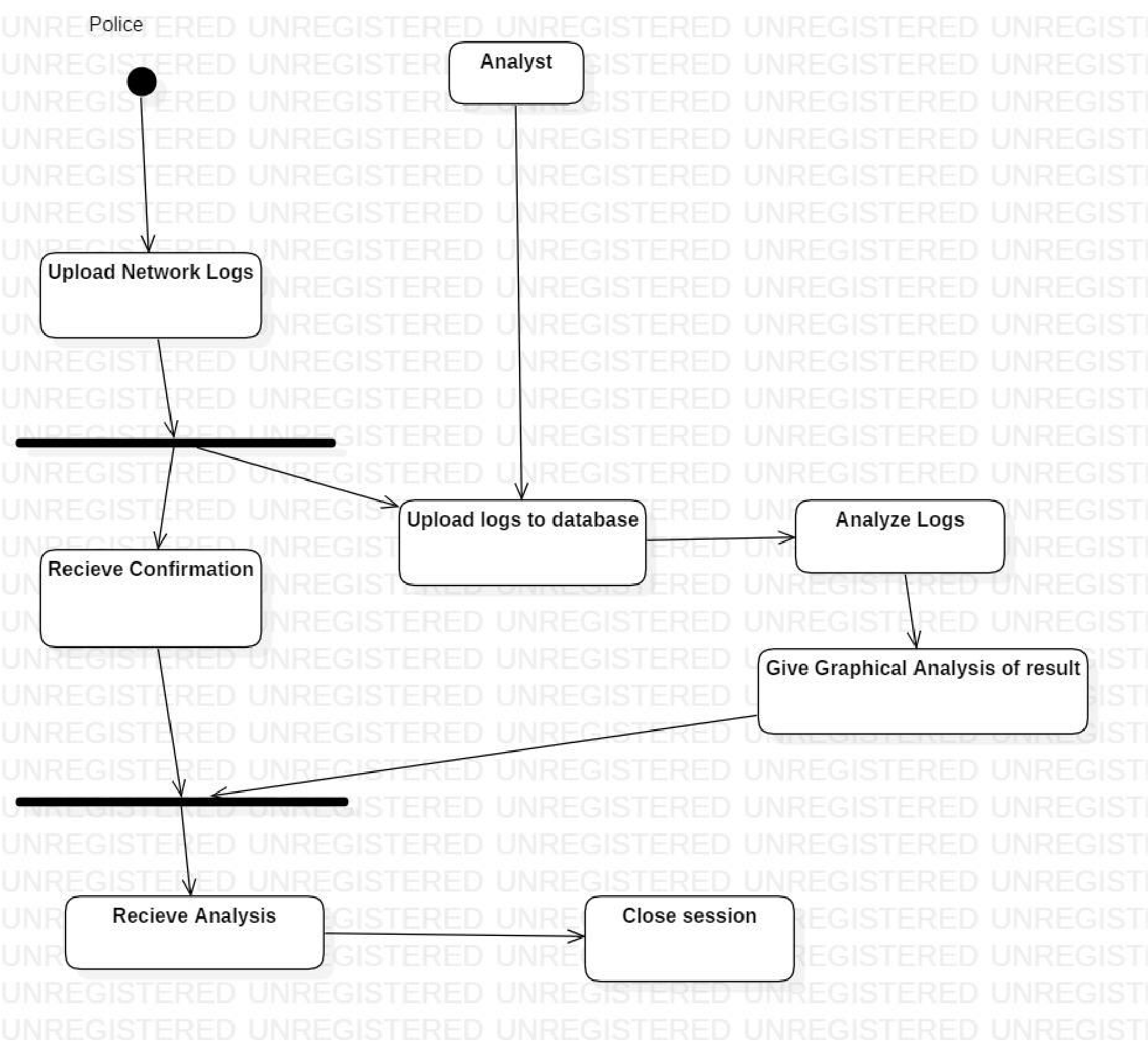


Fig. 3. – Activity Diagram

*B. BACK END*

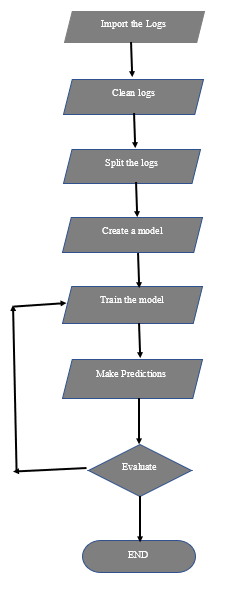


Fig. 4. – 7 steps machine learning flow chart

In this project we are following seven stepped machine learning flow chart which are as followed:

1. Import logs.
2. Clean logs
3. Split logs into training and testing sets.
4. Create a model
5. Train the model
6. Predictions
7. Evaluate

**STEP 1 IMPORT-** For the purpose of the project the we first need the acquired logs are using various tools like packet sniffers. Like the logs used here are network logs for network logs to be captured using wireshark which allows us to view and analyze the network logs. These wireshark logs in .csv format are then manually uploaded to perform machine learning on it.

Here the two logs are mixed too one where the attack is captured and another with normal network traffic.

**STEP 2 CLEANING-** In this step these huge number of logs are cleaned by removing problems of redundancy and empty values. This is done with the use of function like “duplicated” to remove duplication and “fillna(-999)” which replaces the null value in the logs with “-999” value.

**STEP 3 SPLIT-** Here we divide this cleaned data of logs in to two parts, one of 80% and left of 20%.

1. this 80% of data is kept to train our machine learning model.
2. This 20% of the logs are called testing data as these are saved for later to test. According to the majority of data that trained our model earlier will lead the model to itself process this testing data and produce predictions and decision without even programming it to do so.

**STEP 4 CREATE MODEL-** This step involves selecting the machine learning algorithms which will help in finding the results we want from this model.

The algorithms used here to check if a Dos attack occurred or not are as follows:

*1)Logistic regression:* This is a sklearn library’s algorithm, is used to predict the categorical dependent variable’s probability and predict a definite value. This algorithm can be used for various classification problems such as spam detection, we used this in our project to get a value 1/0, yes/no of an attack.

*2)Support vector machine*: This is supervised machine learning algorithm, used to construct a hyperplane in multidimensional space to separate different classes. This algorithm is effective but fails when multidimensional dataset is given to it.

*3)Stocastic gradient descent*: Stochastic gradient descent calculates a derivative from each training data instance and then calculates the update instantly. This helps in finding local and global minimum and maximum. It is useful only when doing "optimization".

*4)k-nearest neighbours*: this is used to solve both regression and classification problems. It is simple and instance-based learning algorithm. It assumes analogous things exist in close proximity. Can be used in nonlinear dataset and it is used to perform classification of various data of the dataset, very useful to group data in our case.

*5)Gaussian naive bayes*: Naïve Bayes used as a probabilistic classifier is a linear classifier of supervised machine learning method and. When processing real-time data with continuous distribution, Naïve Bayes classifier contemplates that the big data is generated through a Gaussian process with normal distribution. Gaussian naïve bayes is its extension with functions like estimation of distributed data.

*6)Decision tree*: this breaks the dataset into smaller subsets eventually resulting in prediction.

*7)Random forest*: This algorithm includes various algorithms of the same type like multiple decision trees, etc resulting in a forest of trees. it can be used for both classification and regression tasks. Very useful in this project.

*8)Adaboost*: adaptive boosting is full form of adaboost classifier is a meta-estimator first this fits a classifier on the original dataset and then additional copies of the classifier are fitted on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. This is extremely good for reducing losses.

*9)Gradient tree boosting*: this algorithm trains the model gradually in sequential manner. This optimizes a user specified cost function, in place of a loss function which usually offers less control and does not essentially correspond to real world applications.

**STEP 5 MODEL TRAINING-** In this step we feed the model we just created with the 80% of training data and 20% of testing data. Which is done by using “Random\_forest.fit(X\_train,Y\_train)” function. The above-mentioned algorithms find the patterns existing in the training data which maps the given data to give out the predictions required and machine learning model is the output that captures these patterns.

**STEP 6 PREDICTION-** In this step we get the prediction that a denial of service attack is occurred in the logs or it’s a normal log based on our model. This is done by using “Random\_Forest.predict(X\_test)” function.

**STEP 7 EVALUATION-** In this step, three factors are determined namely accuracy, confusion matrix and classification report.

Accuracy determine by what percentage is the result accurate which is predicted to be 99% in this case

Confusion matrix is used where we need to get the outliers, since our dataset contain attacks this happen rarely that is why this is very important.

And classification report shows the actual status of the result which are divided into four categories which are true positive, true negative, false positive and false negative.

In this step various graphs are plotted by taking attributes of the logs- Time, Source I.P, destination I.P, Length on X axis and Label on Y axis. Labels is the value introduced by us which is 1 if attack is predicted and 0 if there is no attack.

After evaluation following factors are computed,

*1)*Area under curve for True positive rate-False positive rate

*2)*precision score-(tp/(tp+dp))

*3)*f1 score- 2\*((precision\*recall)/(precision+recall))

If accuracy is less then we train our model again

*C. DATABASE*

For database connectivity we have used MySQLdb. MySQLdb is interface for MySQL that provides python database API [14]. Relational database is deployed to store the logs in .csv format.

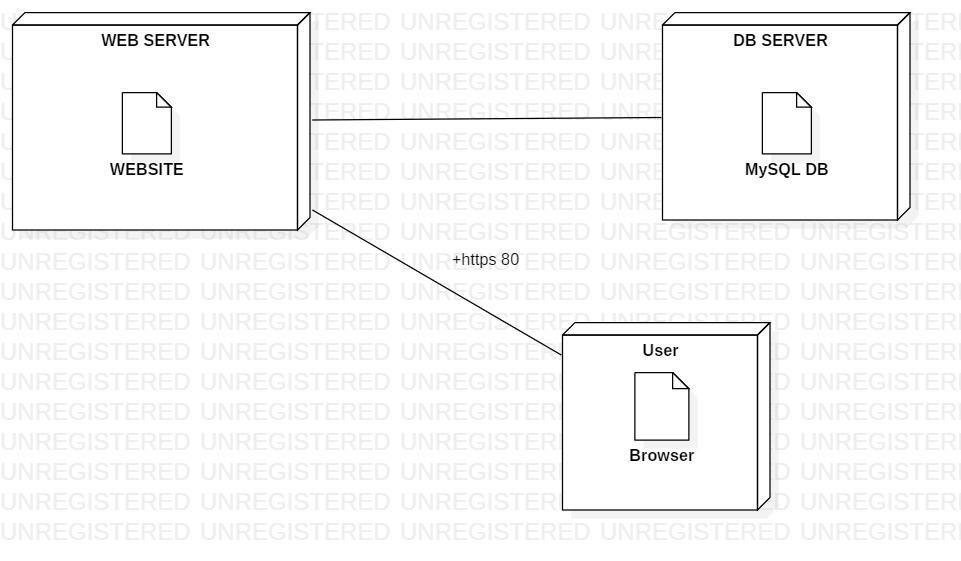


Fig. 5. – Deployment Diagram

In CSV Comma-separated valuesformat refers to text file where data records are denoted by values separated by commas and data record is denoted by each line of the file. In each record comma separated attributes. One or more attributes can be there. In CSV file tabular data is depicted by plain text, therefor every line should have the same number of fields [16].

All log files get uploaded into a database for maintaining data integrity while analyzing and for security purposes.

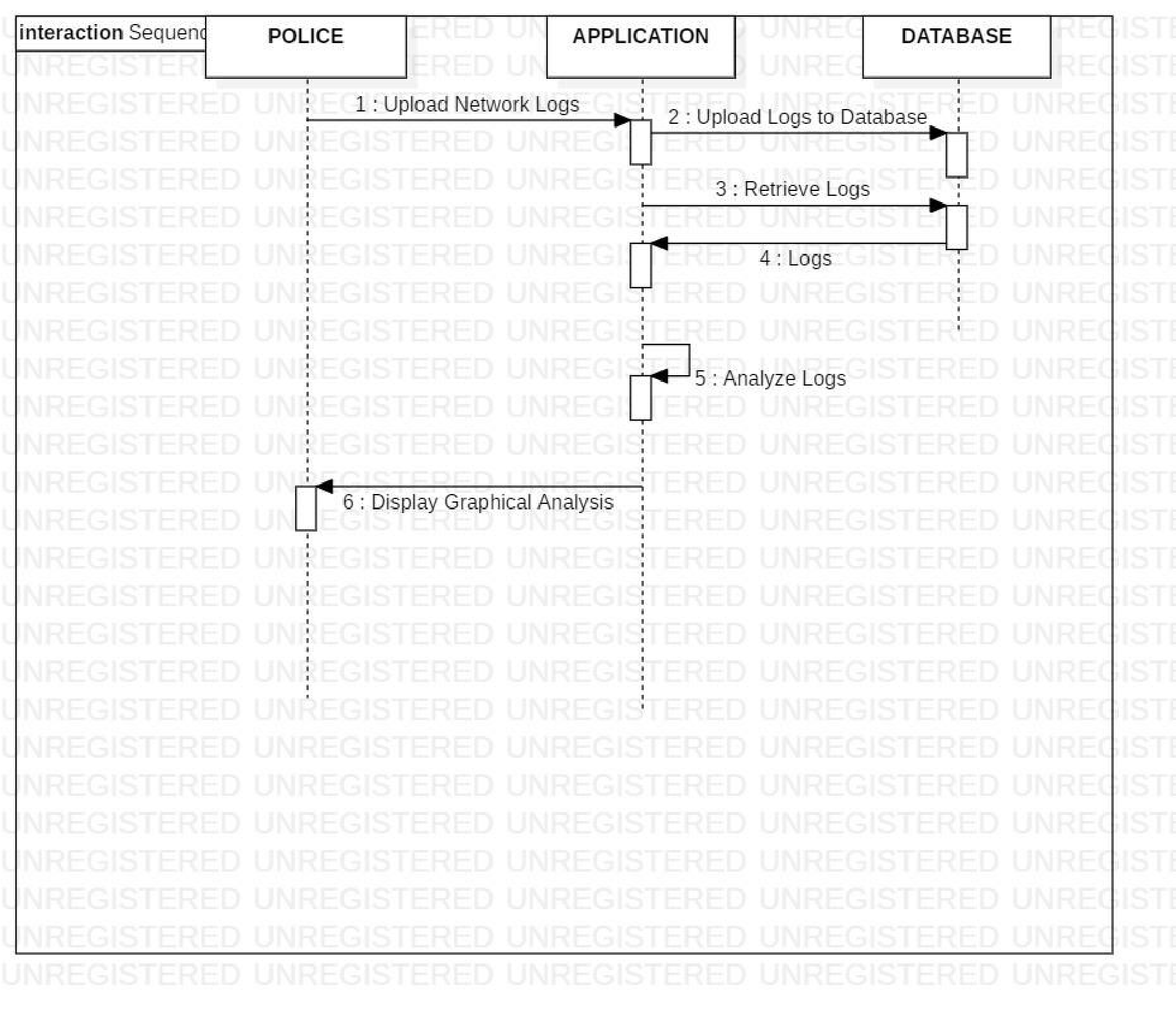


Fig. 6. – Sequence Diagram

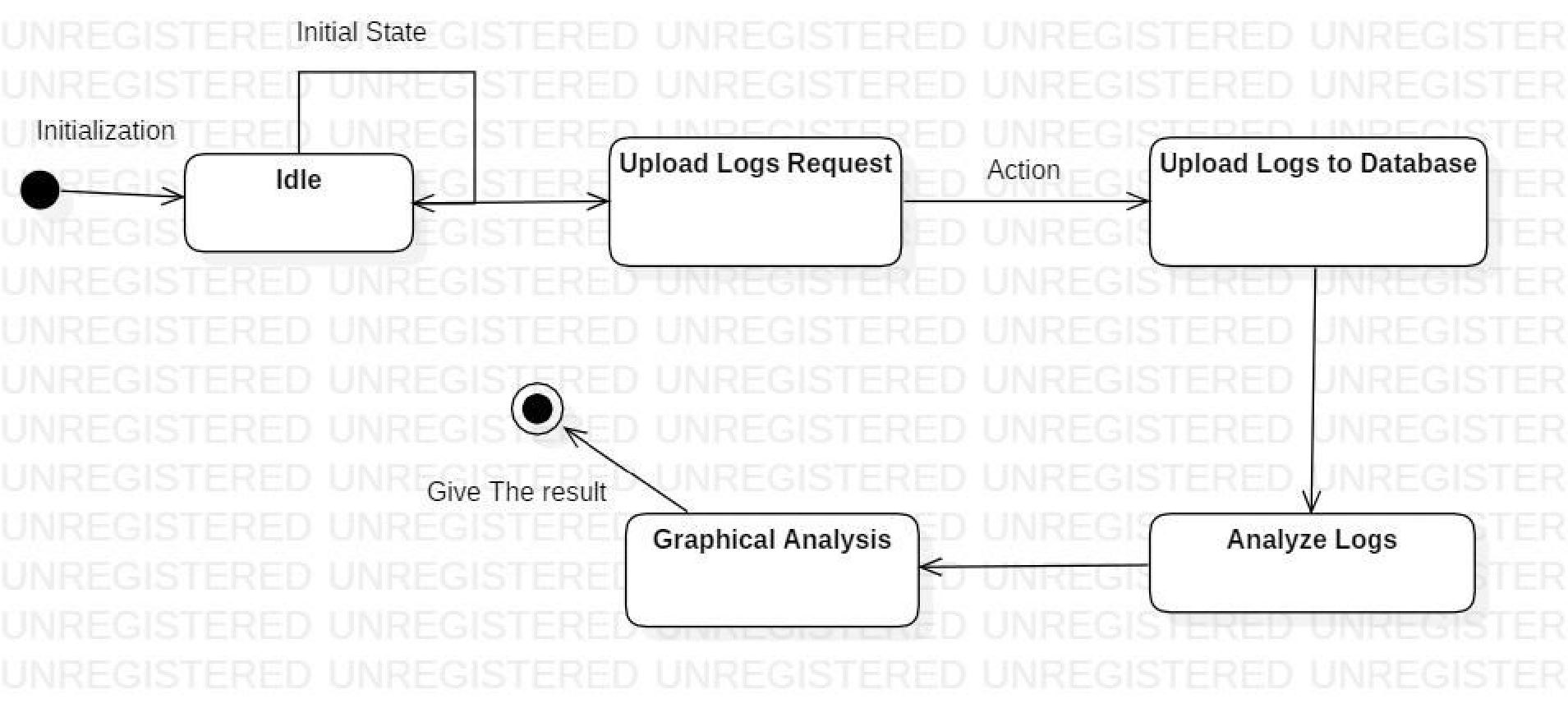


Fig. 7. – State chart Diagram

IV. RESULT

PREDICTION – 1

ACCURACY – 99.98%

V. DISCUSSION

In this big data set are required recommended more than a million rows approximately 10 Gb to get good accuracy and to avoid overfitting or underfitting. Because if the data is small then overfitting occurs in which model remembers the data as it is. And the columns should be enough too because if they are less underfitting occurs.

Data preprocessing is big part of this project because data is huge and the data values need to change from string to numerical data by using different set of algorithms

For good prediction this feature should be added to the project which checks the data and takes the data needed for the prediction and drop the rest data which can lessen the accuracy

Future scope of the project is that this model can by modified to detect other types of cyber-attacks too

VI.CONCLUSION

In this project, several experiments were performed by attacking from other systems to create the attack logs to tested efficiency and the performance of the algorithms used the model. All the tests were done generated dataset. Other attacks are also performed to check the accuracy to detect DOS attacks only. Performance metrics like accuracy rate, precision, false negative, false positive, true negative and true positive are calculated, and false positive and false negative were not accepted. It has been found out that no single machine learning algorithm can detect all the types of attacks efficiently. The model achieved accuracy of 99.98 percent and precision as 1.

REFERENCES

1. Bhavna Dhamarkar, Rajni Ranjan Singh “A review of Cyber Attack Classification Technique based on Data Mining and Neural Network Approach” International Journal of Computer Trends and Technology (IJCTT) – volume 7 number 2– Jan 2014
2. 1V. Hema and 2C. EmilinShyni “DoS attack detection based on Naïve Bayes classifier”

Middle-East Journal of Scientific Research 23 (Sensing, Signal Processing and Security): 398-405, 2015 ISSN 1990-9233

1. Kunal Gupta, Rajni Ranjan Singh “A survey on web application attack detection methods” International journal of advanced research in computer science Volume 8, No.7, July-August 2017
2. Bhavna Dharmakar, Rajni Ranjan Singh “Cyber Attack Classification using improved Ensemble Technique based on Support Vector Machine and Neural Network” International Journal of Computer Applications (0975-8887) Volume 103- No.11, October 2014
3. PyQt5 Reference Guide -https://www.riverbankcomputing.com/static/Docs/PyQt5/.
4. Scikit Learn https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
5. Wireshark - https://www.wireshark.org/docs/wsug\_html\_chunked/
6. Pandas- <https://pandas.pydata.org/docs/>
7. Shailendra Singh, Sanjay Silakari “An Ensemble

Approach for Cyber Attack Detection System: A

Generic Framework” 14th ACIS, IEEE 2013. Pp 79-85

1. Hoa Dinh Nguyen, Qi Cheng “An Efficient Feature)

Selection Method for Distributed Cyber Attack

Detection and Classification” IEEE 2013. pp 1-6.

1. Gradient tree boosting <https://en.wikipedia.org/wiki/Gradient_boosting>
2. <https://en.wikipedia.org/wiki/File:Stachledraht_DDos_Attack.svg>
3. <https://en.wikipedia.org/wiki/Comma-separated_values>
4. <https://en.wikipedia.org/wiki/MySQL>
5. <https://en.wikipedia.org/wiki/Machine_learning>
6. Sergio Alejandro Gómez, Pablo Rubén Fillottrani. "Chapter 19 Materialization of OWL Ontologies from Relational Databases: A Practical Approach", Springer Science and Business Media LLC, 2020
7. Wenhao He, Hongjiao Li, Jinguo Li. "Ensemble Feature Selection for Improving Intrusion Detection Classification Accuracy", Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science - AICS 2019, 2019