National College of Ireland

End-to-end attack detection based on ML and spark

MSc Research Project
Cybersecurity

Ajay Ashok Kumbhar
Student ID: 21138222

School of Computing
National College of Ireland

Supervisor: Dr Arghir-Nicolae Moldovan

National College of Ireland

Project Submission Sheet

School of Computing

Student Name: Ajay Ashok Kumbhar

Student ID: 21138222

Programme: MSc in Cybersecurity **Year:** 2022

Module: ……MSC Research Project……..……..

Supervisor: ……Dr Arghir-Nicolae Moldovan …….……

Submission Due Date: ……15th December 2022……..……..

Project title: …… End-to-end attack detection based on ML and spark……..……..

Word count: ……6801……**Page Count**……22……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Ajay Kumbhar……

Date: ……15th December 2022……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

End to end attack detection using the Machine Learning and spark.

Ajay Ashok Kumbhar
X21138222

Abstract

The main area of concern in the IT environment is the security. There are various attack occurs on the end devices and the networking devices. However, in the current market there are various methodology is present to identify and prevent the attack. Research implemented numerous prevention technique based on the machine learning and the deep learning, however, the accuracy and the prediction rate of some algorithm was good and accurate. Moreover, prediction of attack on the end device is crucial part because most of the attack occurs on the edge devices such as firewall and router. There is so many research is occurred on the intrusion detection system and machine learning is the proposed model from various researcher. Machine learning is the superior and accurately detect or predict the input. To build the machine learning dataset is required and machine learning train by using the available dataset. Apache spark is the huge data processing framework which helps to process the data which is generated in the device. The uses of spark are more beneficial in real time environment where traffic is recorded continuously in the machine.

The aim of this project is to build the end-to-end model to detect the attack. In the proposed system Spark, Random Forest, Decision tree and Binary classification has been implemented. The implemented project is used to detect the attack on the end device and provide the real time notification to the administrator department for the further action. Spark is the most important part of this project which is used to process the larger amount of dataset quickly. Frontend socket is implemented to transfer the captured traffic to centralised server for the prediction of the attack and based on the result of machine learning is generate the log file for specific attacked traffic. Evaluation has been performed for the proposed model and generated values are recorded for the future and to understand the best model. After training and testing we have gathered 97 % accuracy for the Binary classification and 99 percent for the Random Forest and Decision Tree.

Keywords: **K-Nearest Neighbour, Transmission control Protocol, User Datagram Protocol.**

1 Introduction

Nowadays, end devices are main target of the attacker. However, end devices such as servers, computers are the devices which are get impacted by this attack. There are numerous attacks are very famous to lower down the system resources such as SQL injection, DDoS or DoS attack, Cross site scripting attack. During such attack resources were not available to the genuine customer and end devices get hacked or attacked. In the current market various devices available to identify and detect the attack, for instance, IDS/IPS, firewall, web application firewall and more. Identification of attack on the end devices is very important whenever attack occurs from the internal network.
Everybody uses the internet for social activities, banking, e-learning, and other things. Increased internet usage motivates hackers to assault and steal vital data via the internet. DDoS and other types of attacks involve overwhelming the end user's server or devices in order to cut off the real user. However, faked DDoS attacks are highly challenging to spot since the attackers imitate the source IP of the packet or utilize numerous bot machines to flood the internet with traffic for a given destination, resulting in system failure and loss of service. For simple network devices, differentiating between authentic and fraudulent traffic is a complicated process.

AI has potentiality to exhibit human cognitive abilities that are connected to the human mind. The study of intelligent agents—systems that can comprehend their surroundings and act to achieve their objectives—is called artificial intelligence (AI). Speech recognition, machine learning, are examples of common artificial intelligence approaches. In numerous studies, machine learning is the most often used methodology to predict assault.

As per the report (Anon., 2022) it is recorded that 28% raise in the global attack in the third quarter of 2022 which is higher than the previous year for the same quarter. The report also mentioned that weekly average rate is climbed above 1,130 per organization. Education or Research industries was main target in this quarter which showed 2,148 attack per organization. Along with this, healthcare system was mainly targeted for ransomware.

The Intrusion detection system is the software or hardware device which continuously monitor the network activity for the malware. Basis on this unexpected behaviour it notifies to administrator team or block the traffic as per the policy configured on the device. In IPS/IDS there are various types are available for example, Signature based inspection, Anomaly detection, policy-based detection. These are some configuration modules are present in the IDS/IPS. (Amanoul, et al., 2021) According to number of intrusion system provides the false alarm in a high amount along with this smaller number of warnings for the low-threat.

As per the researcher (Amanoul, et al., 2021) they mentioned that machine learning is the intelligent technology which could be used to identify the attack basis on the huge amount of dataset available on the market. Depend on the great amount of dataset machine learning model can accomplish better sensitivity.

This paper is examining the potentiality of the deep learning model to detect the attack on the end device. However, this model used the latest dataset which is available publicly in the market. As per the paper (Damaševičius, et al., 2020) this is real-world network flow dataset was introduced to keep up-to date development. The paper mentioned the comparison of old

dataset and Newley launch dataset. The LITNET 2020 dataset has various features which could help to detect attack on the network system. The dataset was captured and stored in the SQL server with the help of the NetFlow protocol and generated required additional features for the dataset. The proposed dataset in collection of more than 10-month period records.

The model used the random forest deep learning methodology to detect and predict the attack and along with this spark library is implemented to process huge amount of dataset when it captured from the multiple end system.

The below points illustrate the important factor and main aim of the project.

- This paper is unfamiliar with others paper because of it is end-to-end attack mitigation by handling the huge amount of dataset. For to handle large dataset we used spark framework.
- This paper includes the recently available dataset which is LITNET 2020.

There are numerous research papers in the literature that discuss mechanisms for detecting and stopping fraud activity. There are three distinct stages in the paper below.

A. structured process is used to collect more focused research data for the Intrusion prevention study.
B. A recent paper is used to present the most recent information in the field of DDoS identification, which will improve the defences against fraudulent activity.
C. On proposed research, implemented the model and evaluated the performance of the model and recorded the important factor in this paper.

1.1 Research Question

By implementing the network security system, how machine learning based mode could be effective on the various intrusion which occur on the network? How we can achieve the better accuracy by balancing the dataset with different classifiers? How can spark and machine learning based model can be used implement protection tactics against the intrusion attack?

2 Related Work

Detecting end to end attack is the crucial part of the cyber security, because most of the attacks are internal. However, there are so many research occurred on the identification of attack and prevention based on the different models and machine learning models. Furthermore, machine learning and the deep learning models are the widely used in the nowadays.

The paper (Zhang, et al., 2016) worked over the Hadoop and machine learning model. Hadoop is the commonly used big data model and this is the very old model. Moreover, the paper also focused on the HBASE and the bloom filter. Hadoop framework is openly available in the market and it is required to store and execute the high number of values of dataset. Hadoop has some main modules which are HDFS, MapReduce, YARN and Hadoop common. However, as per the research paper (Zhang, et al., 2016) they have used the MapReduce to identify and predict the attack. The paper also concluded that uses some algorithm such as CUSUM will identify attack more accurately and efficiently. The goal of this paper to identify and predict DDoS with the spoofed IP address.

(Priya, et al., 2020) The purpose of this article is to examine several DDoS traffic types, including ICMP, TCP/UDP flood, and others. However, according to earlier research, some algorithms were only usable on a certain type of traffic, a limitation that has been overcome in this work. The numerous algorithms utilized in the previous paper were the subject of research. Additionally, for accuracy improvements, they updated a number of algorithms, including Random Forest, KNN, and Naive Bayesian.

As per the research from paper (Lakshmanarao, et al., 2022) it is observed that they have used the two different datasets from for the prediction of the attack. As compare this paper with the previous one it clearly indicates that the focus of this research on the various cyber-attack. Also, the paper has used different dataset which could provide the better result than the single set dataset. The research mentioned that they have used various ML algorithms to predict the attack which are K-NN, Logistic regression and feed forward. However, research concluded the accuracy of 88 and 99.9 percent for these two datasets. The dataset used on this research was CICIDS-2017 which is very old dataset. This dataset has included various attack and normal traffic as well. The attack included in this dataset was Brute force, DoS and DDoS, XSS, SQL injection and more. The dataset is publicly available in the market and it has also included the original raw files of the captured traffic during the attack and normal flow.

(Amanoul, et al., 2021) The paper worked over the machine learning to make improvement in the intrusion detection system. The paper included the various machine learning algorithm and their output in the accuracy. The implementation occurred over the dataset KDD Cup 99 which is very old dataset. The researcher used various deep learning model to predict the attack and mentioned the accuracy of the model. As per the report is declare that 98 percent accuracy observed from the BayesNet, however, 99.98 and 99.35 from Random Forest and Neural network, respectively. The paper concluded the accuracy of the different model which used in this research.

The author (Pérez-Díaz, et al., 2020) concentrates on the undetectable low-rate traffic. Additionally, research is being done to protect the SDN network against low-rate DDoS traffic; for this purpose, machine learning and deep learning mechanisms have been incorporated. The article investigates SDN architecture, but it may also be applied to conventional systems by applying machine learning techniques like Random Forest, SVM, REPTree, and MLP. The Hadoop and HBase approach were used to gather a large amount of data, including CPU use, TCP connection and count packet size. The neural network forecast was tested over the course of three 9-day periods, including normal, offensive, and active days. However, additional crucial aspects that will be covered in this study must be worked on in order to forecast faked IP addresses.

With regards to end-to-end identification and prediction of attack it is very important to capture and store the data centrally. However, because of the number of end device there is possibility of the huge amount of dataset and to work smoothly on those datasets require the superior big data process framework. Moreover, in our case spark will fill the place for huge data processing and machine learning model will run to identify and prevent the attack which will occur on the network.

The below subsection provides the information regarding the IDS system and the use of machine learning model in the intrusion detection system.

2.1 Intrusion detection basis on the anomaly detection:

The research conducted from (Zhao, et al., 2015) it is discovered that, they used the big data processing algorithm such as Hadoop, Kafka, and Apache storm with for prediction they used some machine learning algorithms. However, they have used the NetFlow database which was collected in Missouri-Kansas university. For the streaming the big data they used Kafka and after that for the processing it is passed through MapReduce. To produce the output machine learning has used different features which are source port and IP, destination port and IP. The researcher used three algorithms in machine learning models such as SVM, Naïve Bayesian and the decision tree. The author mentioned the 91 percent accuracy from this model.

As per the research occurred from authors (Cui & He, 2016) it is observed that, Hadoop was the major big data process model in that era. However, nowadays there are many new other big data processing models are available which are very good in real-time data processing. Hadoop is the oldest and use the batch processing functionality. Moreover, the authors implemented the algorithms which are good in accuracy which are SVM, naïve Bayes and Decision tree. The authors mentioned that the observed the more than 90% accuracy from this model.

2.2 Intrusion detection basis on signature based:

Signature based detection is another procedure is implemented in intrusion detection which match the specific data with the previously occurred attack.

As per the (Jongsuebsuk, et al., 2013) it can observe that they have used the Hadoop and the fuzzy logic algorithm for prediction of the attack on the network. The researched trained the model with the help of the known dataset which is called KDD99. MapReduce was implemented to remove the unrequired data with the help of conjunctive rule. However, the output of the MapReduce is forwarded to the decision-making system to predict if there is attack or not. Author mentioned that, 2 to 3 seconds were taken to detect the attack using online dataset and few seconds to detect KDD99 dataset.

The paper (Malek, et al., 2020) included the pattern and signature-based intrusion detection system. In this detection system they have focused on the user activity after log-in to the device. Researcher proposed the combined model of Statistical and Pattern based intrusion detection model. Author used Zakiya dataset which contained various parameters of user interaction with the system. Jess is used to produce the rules for pattern-based intrusion. The result generated from this rule are compared with the database of Zakiya. The researcher recorded the 75 % accuracy from this model.

2.3 Intrusion detection with the Machine Learning:

Nowadays, Artificial intelligence is superior methodology to predict the result basis on the database. As per the research conducted from (Saxe & Berlin, 2015) they have consumed approx. 400k software binaries to build deep learning system. There was three layers in the model which are training layer, distribution of the data and final classifier layer. The researcher recorded the 95 % accuracy but cleaning of data is not present in the model.

The author (Al-Maksousy, et al., 2018) mentioned that the dividing the system into the two different part such as detection and analysis could improve the accuracy of the detection. The author used the following classification methods to detect the attack in the network which are random forest, SVM, Naïve Bayes and deep neural network.

With regards to the below proposed model, we can achieve the high accuracy with the help of random forest and spark could be useful to handle huge dataset when it generates from the multiple end device.

3 Research Gap:

Research paper are cited in the literature review which will helpful for the future research on this project. However, after the research we understood that there is still gap is present in the research. Most of the time attack occurs on the edge of the devices and end devices as well. The research recorded the great accuracy and author also used the better Machine learning model to acquire great performance. Usually, to predict the model required the huge amount of dataset and network need to store data continuously. When the dataset is huge it takes time to process, hence high data processing model is required. However, most of the research has used old dataset which could provide the poor accuracy and the model will not be that much useful. To overcome this issue, we need to build end to end model with high data computing platform which is Apache spark and latest dataset to train the model.

Below table [1] display the summary report of reviewed paper along with various information.

Research Paper	**Dataset**	**Algorithm**	**Proposal**
(Zhang, et al., 2016)	NA	HBASE, TCP2HC/UDP2HC, CUSUM	Research build Hadoop based system
(Lakshmanarao, et al., 2022)	Kaggle, CICIDS-2017	K-NN, Decision tree, Feed Forward,	Proposed the model with different dataset.
(Amanoul, et al., 2021)	KDD Cup 99	Bayes Net, Random Forest, Neural Network, RNN, LSTM	Proposed system for the IPS.
(Zhao, et al., 2015)	UMKC campus	Apache Hadoop, Apache Kafka, Storm	Proposed real time model using Apache spark
(Cui & He, 2016)	KDD CUP 99	naïve Bayes, decision tree and SVM, MapReduce	The researcher focus was to build model foe huge dataset
(Jongsuebsuk, et al., 2013)	KDD99	Fuzzy	Proposed the real time model using fuzzy
(Priya, et al., 2020)	Manually generated	Machine Learning KNN, RF, NB	Proposed machine learning model.
(Pérez-Díaz, et al., 2020)	CIC DoS	Machine Learning	Random Tree, REP Tree, Multi-Layer

			Perception and support vector machine, Random Forest,

Table [1]. Summary of reviewed paper.

4 Methodology

As per the research we observed that there is various mechanism is available for the different attack on the network. However, most of the research is combined the multiple machine learning algorithm to identify the accuracy of the model. Moreover, the model will provide the end-to-end attack detection base on the latest dataset is available in the market. This model could quickly work over the huge amount of dataset because of the uses of spark libraries.

4.1 Dataset selection:

To build the machine learning model required the specific set of datasets and the required features to predict the model is there any attack or not. After the literature review, we observed the various dataset is available in internet and few was the most commonly used in the previous research. We mainly focused on the latest dataset which might be contained the good amount of traffic and mostly with the latest features. LITNET 2020 is the latest dataset was taken in this model to predict the attack. (Damaševičius, et al., 2020) The LITNET 2020 is the real-world dataset was stored with the help of NetFlow service. The dataset has contained the various features which could be useful for various network-based model. The dataset collected from the University KTU 1 and KTU 2. In this network environment NetFlow Server was hosted which stored the NetFlow traffic in the SQL server with the help NetFlow service enabled on the KTU firewall. Figure [1] shows the process flow diagram for the dataset. The flow consists of Data selection, analysis of the data after that the main component which is balancing the data and processing. Eventually, dataset could be split for training and testing for the specific model.

We split the dataset in 60 and 40 partition for training and testing the model. The splitting is required to not to overfit the model. (Gillis, n.d.)

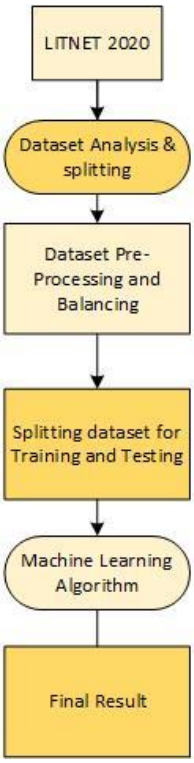

Figure [1]: Model flow diagram.

4.2 Dataset preparation and Balancing:

Usually, the dataset which is available in the online platform will be unbalanced and more processing of dataset will be required before training the machine. Moreover, as per the model need to identify the required features which will help to predict the attack or result for that specific model. LITNET dataset has contained the various fields features. Before putting the dataset, we had required to pre-process the dataset for better output. With the help of python commands, we have identified the some of the fields could be useful for the model. However, next step was to identify which model will be best suit for those data. Data visualisation is performed to find the correlation, trends and pattern of the data which can't be view when dataset is in CSV format. The next step was to recognise the dataset is balanced or not. As per the dataset view, we observed that there were some null values which is replaced, however, some fields were included categorical values which is replaced to integer. In our database there is multiple columns are present and to perform the data standardization we used **fit_transform** method to do.

The above method is important while training the model, the **fit_transform** could be utilized in the training data so we can learn the scaling parameters of the data. However, it is need to use to scale the training data. (Goyal, 2021)

After the replacing categorical values, we had performed data balancing, because as per the data visualisation we observed the data was unbalanced. Using the labelling model which is label encoder we have labelled the normal and attack traffic to keep data presentable.

4.3 Selection of features:

Most important factor in the Machine Learning is to find the important and useful information for the model. However, feature selection is the important part of the model and to identify the features we have balanced the data with the KNN classifier and exported the clean data. The clean data is having the features with higher values in those columns. As per the output we recorded the total 19 columns is having the great features. Below table [2] is the list of the features which was contained the better values in the columns. (Damaševičius, et al., 2020)

Features	Description
sp	Source Port
_dp	Destination Port
pr	Types of protocol such as TCP, UDP.
flag1	_TCP flag 1
_flag2	TCP flag 2
flag3	_TCP flag 3
_flag4	TCP flag 4
_flag5	_TCP flag 5
flag6	TCP flag 6
ipkt	_IP packet
ibyt	IP bytes
opkt	Output packet
obyt	Output byte
_in	Input traffic
out	Output traffic
attack_t	Types of attack
attack_a	Normal or Attacked traffic

Table [2]. Key features of database.

After the viewing dataset we found some important features which could be most useful for the future prediction process. Below is the list of features which I have chosen to implement the model.

5 Design Specification

To implement the model, Windows 10 and Jupyter notebook has been used. Jupyter notebook has been installed in the windows 10 to implement the code. For the large amount of data, we required the good amount of RAM and processing power, however, while training the model machine also utilise high CPU while processing. (Singh, n.d.)
Desktop Specification:

CPU: Intel i7-11370H @ 3.30GHz
RAM: 24 GB DDR4
GPU: External GPU GEFORCE with 4 GB RAM
SSD: 1.5 TB SSD

(Anon., n.d.)Whenever we need to work with dataset, we have to clean the dataset for the future process of model. However, as per the dataset I have performed the data cleaning and for that NumPy and pandas has been used. (Devara, n.d.)

5.1 Model design

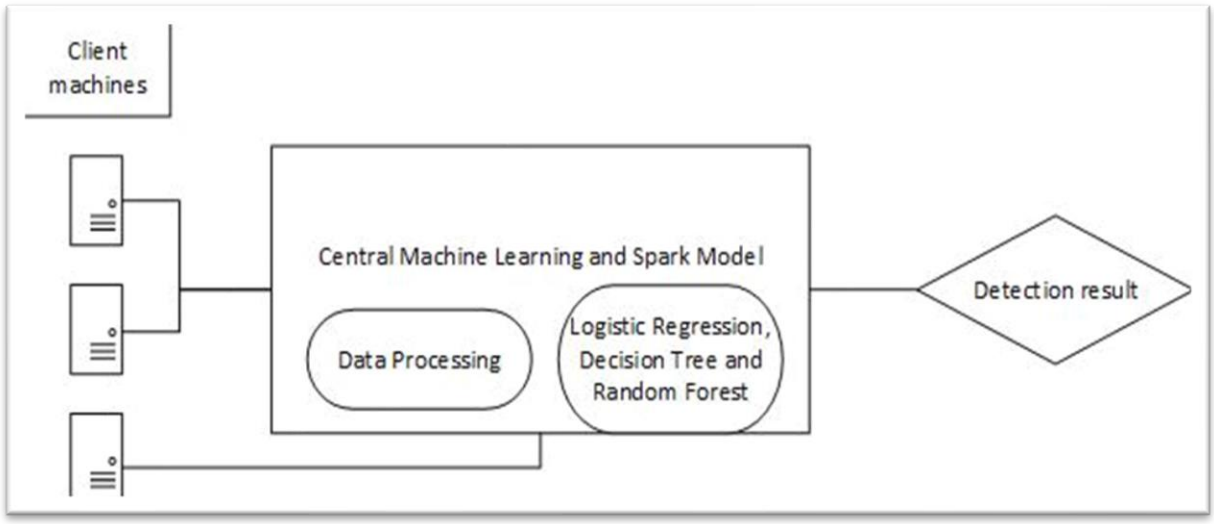

Figure [2]: Model Design.

The above figure [2] is the proposed model for the end-to-end attack detection using the Machine learning and the spark model. Python client has been installed on the end machine to transfer the packet from the client machine to the centralised proposed model. Python client file is the just script which can be run in the multiple client machine to connect to the server machine. However, centralised model is the combination of the Machine learning and the spark model which will process the data quickly and detect if there is attack occurred or not. Pyspark is the library which will connect to the Apache spark using the python. However, multiple machine learning model has been placed to detect the attack with the better accuracy and the performance. These introduced multiple machine learning model will execute to detect the future attack on the received data from the client machine. The output of the model is displayed

at the end of the model. PySpark is the API of the Python for the Apache spark and it is the computing framework. Logistic regression, decision tree and Random Forest will use to predict the attack basis on the trained model.

5.1.1. Reason to select Decision tree:

In case of decision tree, it is a good classifier but it also overfit the model because the model store the training the data and perform the testing on the test which is very nearest to the training values. Random Forest is the key solution to overcome this issue.

We can rely on random forests, which develop an individual prediction method using a collection of decision trees, to solve the problem of unbalanced values and overfitting. The method used in random forests is to train every decision tree with a variable split of training data, resulting in hundreds of decision trees. The parameter estimates that result is the average/voting value of all the forecasts. (Duggal, 2022)

5.1.2. Uses of Binary Classification:

Binary classification the classification methodology which could be useful when in the dataset there is only two values would be present. There are many datasets we use to predict, in that dataset there are chances of values like 1 or 0. However, for such values binary classification will be the better classification method can be used. In our proposal we have LITNET dataset which has different types of attack, features and the most important value which is abnormal or normal traffic. Abnormal and normal traffic might have only two values which is 1 or 0 and due this reason binary classification has been implemented in this model.

5.1.3. Reason to select spark Apache Spark:

In our proposed model we are using the Pyspark and the machine learning algorithm to predict the attack on the edge devices. Whenever, edge devices come in the picture then it required high amount of data processing mechanism. Because implementing end device model we have to take care of huge dataset. Client will share the stored dataset into the centralised server and centralised model will predict if there is attack occurred or not. During the process there are chances to deal with the huge dataset and spark is framework which helps to process the dataset quickly. Pyspark in the library which provides the interface to the Apache spark in python. Pyspark has included many features such as dataframe, Machine Learning and streaming. With the help of the pyspark we could run application on the multiple nodes. The main benefit of the Pyspark in the real time environment where huge number of datasets generate continuously. (Gairola, n.d.)

5.1.4. Uses of Random Forest:

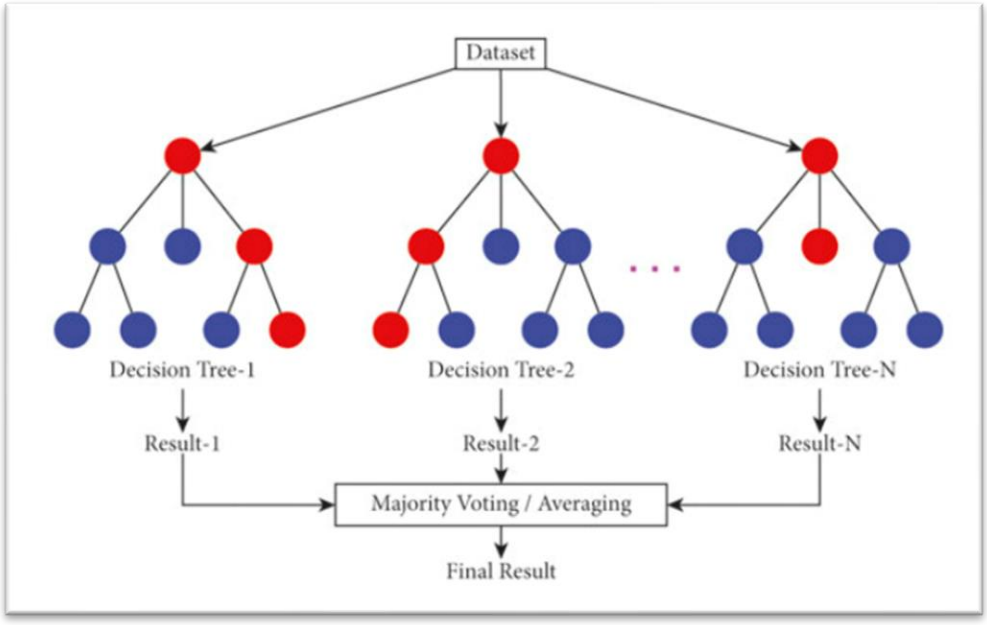

Figure [3]. Random Forest workflow.

Above figure [3] is the sample workflow model of Random Forest. In machine learning random forest is widely use to learn. The "forest" that it produces is made up of a variety of decision branches, which are sometimes taught using the "bagging" method. The main premise of the base classifier is that mixing learning strategies enhances the result. The Random Forest is very useful in the two main factor which are classification and for the regression problem. Also, the previously mentioned issue in the decision tree which is overfitting could be overcome using the random forest. (Mohamed, 2022)

6 Implementation:

To build the model required the specific libraires and the programs to work. However, below is the process which carried to build the Machine learning model with the PySpark.

6.1 Data Preparation:

The deploy the model need to verify the dataset and the required features which could provide better performance after deploying the model. However, in the initial process I browsed the latest dataset which was available on the online platform. After the reviewing the dataset, I understood that the dataset has number of features and values. The dataset was separated into the multiple separate file accordance with the type of an attack. I performed the data visualization to clearly understand how much amount of attacking and normal traffic is present on the chosen dataset. Figure [4] show the count of the normal and the abnormal traffic.

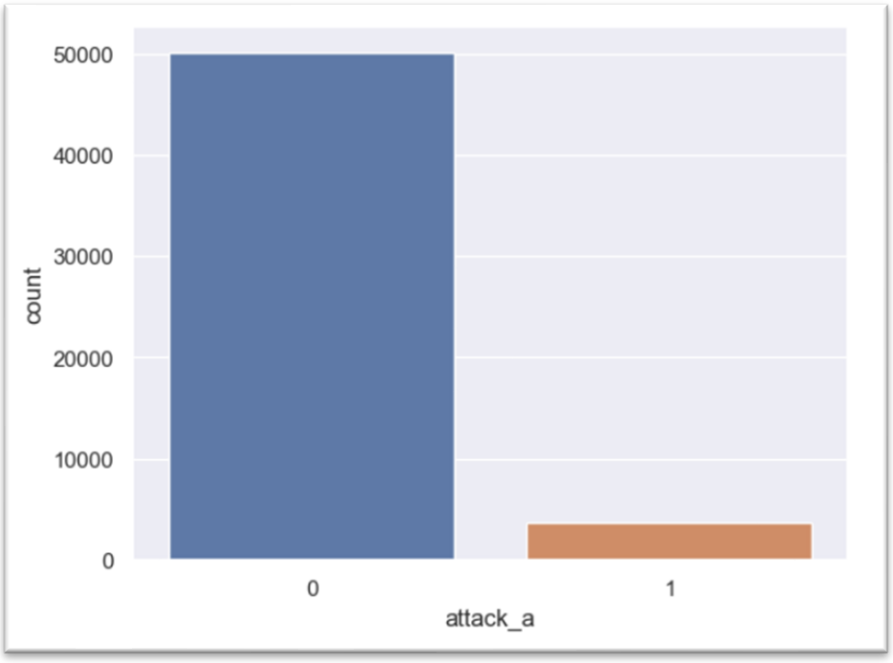

Figure [4]. Normal/Abnormal traffic chart.

The visualisation graph provided the better understanding of the graph. Moreover, for the sub category of the attacked traffic I also performed the visualisation which also recorded the traffic is balanced or unbalanced. Figure [5] is the representation of the sub categorised attack

The figure [5] recorded the sub categorised output of our database. After performing the visualisation of dataset, it is discovered the dataset is not balanced and need to perform the balancing of the data. The graph shows number of normal traffic is huge and abnormal traffic is very less for the other sub categorised dataset.

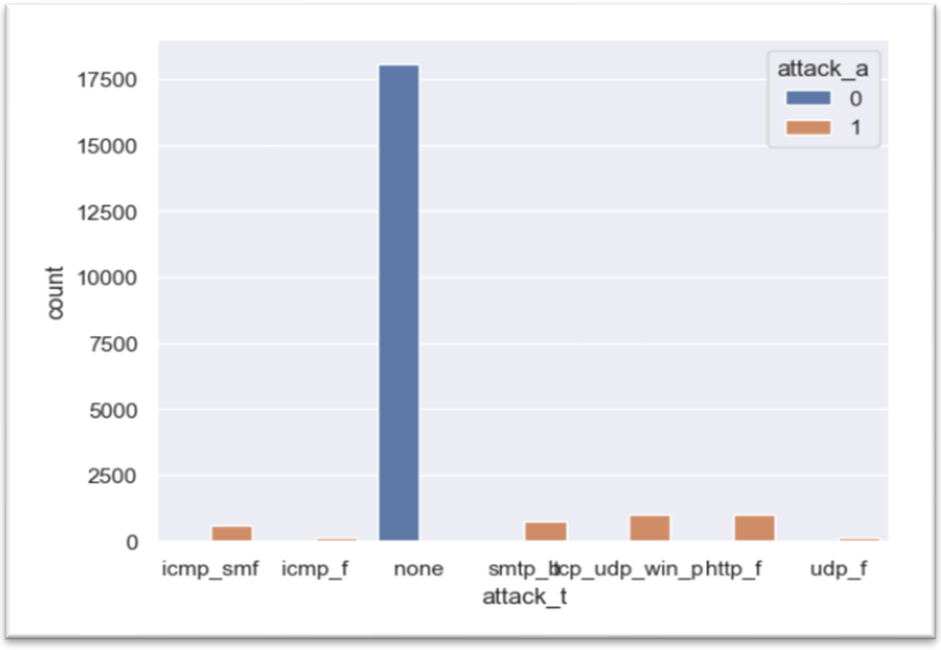

Figure [5]. Sub categorised attack.

6.2Architecture diagram:

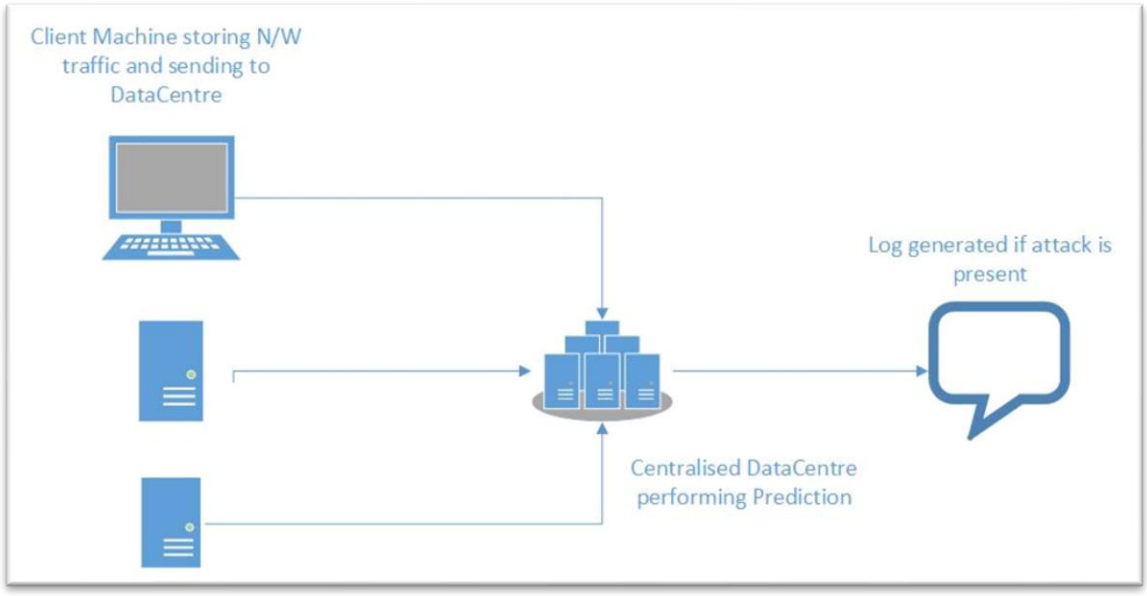

Figure [6]. Network Architecture.

Diagram [6] is the architecture of the proposed model. In this model client machine will run the python file which is included the script. Client machine will continuously store the traffic and will transfer this stored traffic to the centralised server. Connection between server and client, I have created the socket connection on the server side which is available on the port number 5002. The port number 5002 is open for everyone. However, with the help of the IP address and the port number client machine will transfer the captured packet to the central server. Central server is having the three model which are random forest, decision tree and random forest. The proposed and trained system will predict the traffic which is send from the client machine and notify to the administrator team. In the proposed system we have implemented the log file which is storing the log entry of the detect attack. The generated log file will help to block the traffic of the specific client.

Spark is library which has been used in the centralised server to process the generated traffic in the client machine. Spark is the distributed framework which help to process the huge data quickly. Because of this methodology spark has been used here. (Gairola, n.d.) Proposed machine learning algorithm will predict attack. Random forest has been achieved the higher accuracy after training the model.

To implement the model has been used local system where spark service was running and required application such as Visual studio or Jupyter Notebook is used. Using this program, model is implemented and python script has been installed on the client machine to send the packet. In this case we could use multiple clients to share their stored traffic. Using socket connection file get transferred and trained model is detected the attack. We can also host the trained model on the cloud platform for better exposure or we could host on the enterprise server to detect client traffic which is getting affected my attacker or not.

7 Evaluation:

The goal of the study is to ascertain the outcomes of using the Random Forest, Logistic regression and algorithms Decision Tree. However, to evaluate the dataset. We were able to assess each machine learning strategy separately on the data using the dataset we used. The most effective method for protecting end devices would be chosen after considering the accuracy, false positives, true negatives, and positives. Below is the understanding of the confusion matrix and the other assessed result in Model. (Mishra, 2018)

Accuracy: Using the created model, the records may be properly categorized. To determine accuracy, apply the equation below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

F-Measure: It is calculated using the precision and recall harmonic means.

$$\text{F Measure} = 22\left(\frac{Precision \times Recall}{Precision+Recall}\right)$$

Recall: The proportion of values which are positive among those that were correctly identified. The Detection Rate is another name for this.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Precision: This number represents the proportion of attacks that were correctly predicted across all samples.

$$\text{Precision}=\frac{TP}{TP+FP}$$

7.1 Case Study 1: Binary Classification

After the training the model, performed the testing with the remaining dataset. However, after performing the testing we have calculated the accuracy, F1 score, Confusion matrix of the model which help to understand the machine. As per the recorded value of logistic regression it is recorded that total 97% accuracy for the tested dataset. Along with this other required parameter has been also mentioned in paper.

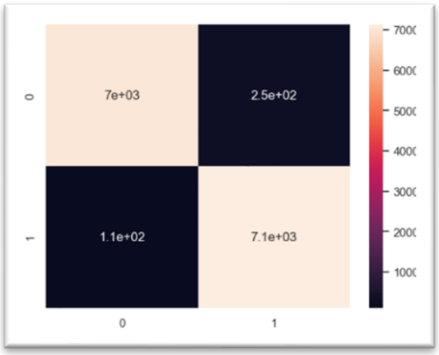

Figure [7] Confusion Matrix.

The figure [7] shows the confusion matrix of the true and probability values. The variables which are (0, 1) are outcomes and index 0 shows the probability and the next one (1) is true. The numeric calculation of above confusion matrix is mentioned in the table [3].

	True	Probability
True	7000	250
Probability	110	7100

Table [3] Numeric Confusion Matrix.

7.2 Case Study 2: Decision Tree

I have trained the dataset with the other model as well to asses which model will provide better accuracy. The decision tree model has been implemented to train the data and after training the model with the dataset. Calculated the required parameters and the accuracy of the model is 99 percent.

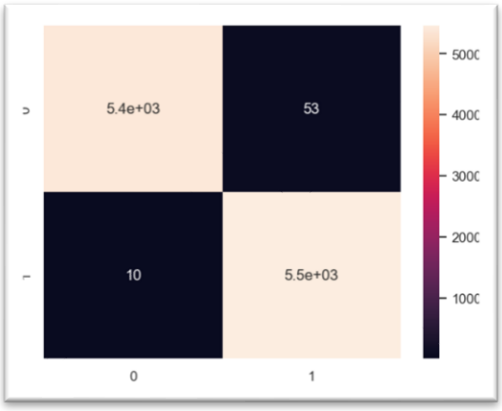

Figure [8] Confusion Matrix.

The above figure [8] is the confusion matrix which is received from the decision tree model. However below table [4] shows the numeric calculation of the confusion matrix.

	True	Probability
True	5400	53
Probability	10	5500

Table [4] Numeric Confusion Matrix.

7.3 Case Study 3: Random Forest

The random forest is the superior learning algorithm which could provide the better accuracy of the model. However, after performing the random forest we have gathered the below parameters. Test data has been transferred to the Random Forest model and evaluated result of the algorithm. After testing, 99 percent accuracy has been gathered from the Random Forest model. Below is the recorded graph of the model.

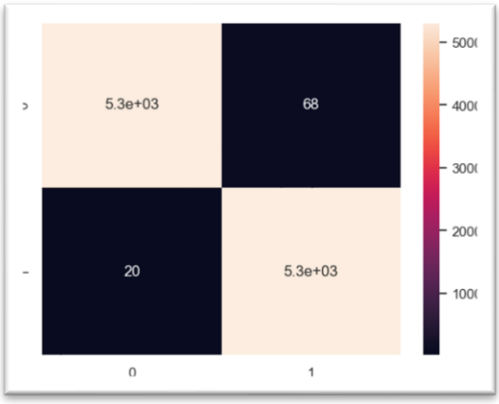

Figure [9] Confusion Matrix.

The figure [9] is the random forest model confusion matrix and it is the best model with accuracy. However, below table [5] is the true and probability values of confusion matrix in the numeric form.

	True	Probability
True	5300	68
Probability	20	5300

Table [5] Numeric Confusion Matrix.

8 Discussion:

After successful implementation of the model, we performed the testing of the model. To test the model, we ran the python script file on the client machine and socket session was open on the server where Machine learning model was also deployed. The asses the model we had use same machine as client and server. Using CMD we ran client machine and Jupyter notebook was used as a server. The socket session was open with the port 5002 and using cmd to the server IP which is also same IP of client and server due to both client and server was on same platform. With the help of socket file is transferred and performed the testing. After the testing we achieved the result with the positive output which means machine quickly identified the attacking traffic and normal traffic.
With the help of all result of confusion matrix for individual model we have calculated the other important factors such as precision value, recall and F1 score. Below is the summarized table of all these values.

The below table [6] shows the accuracy of the model which could be best fit model for intrusion detection system. However, socket session confirmed that this the best model to use in end-to-end detection of intrusion attack on the network system. Moreover, implementation of spark is best framework while working in the large dataset.

Metrics	Binary classification	Decision Tree	Random Forest
Accuracy	97	99	99
Recall	0.96	0.99	0.98
Precision	0.98	0.99	0.99
F1 Score	0.97	0.99	0.99

Table [6]. Calculated values/metrics of different models.

9 Conclusion and Future work:

The main goal of this research is to overcome the drawback of the Intrusion system. Nowadays, machine learning is the highly usable model which could predict accordance our dataset. As per aim, I have implemented the machine which using spark and the machine learning model which predicating the attack easily. The accuracy of the model with different algorithm recorded that this model is complete the objective of our research, however, use of socket session helped to detect the attack on the multiple end device. However, during the implementation many errors are discovered and resolved using the open-source platform. References to the resolution are mentioned. While working over this proposed model, I gathered and learned very useful knowledge which could help for future proposal as well.
After viewing the output and assessing the model. I understood model could be implemented on the cloud platform as well and model can be used for real time environment also. Spark is better in real time environment which also manage huge traffic, hence implementing in real time environment by changing some code could be work. Moreover, by scheduling the job at client side can transfer captured traffic to central server for prediction.

References

Al-Maksousy, H. H., Weigle, M. C. & Wang, C., 2018. *NIDS: Neural Network based Intrusion Detection System.* s.l., s.n., pp. 1-6.

Amanoul, S. V., Abdulazeez, A. M., Zeebare, D. Q. & Ahmed, F. Y. H., 2021. *Intrusion Detection Systems Based on Machine Learning Algorithms.* s.l., s.n., pp. 282-287.

Anon., 2022. *Checkpoint.* [Online]
Available at: https://blog.checkpoint.com/2022/10/26/third-quarter-of-2022-reveals-increase-in-cyberattacks/

Anon., n.d. *activestate.* [Online]
Available at: https://www.activestate.com/resources/quick-reads/what-is-numpy-used-for-in-python/
[Accessed 31 1 2023].

Cui, B. & He, S., 2016. *Anomaly Detection Model Based on Hadoop Platform and Weka Interface.* s.l., s.n., pp. 84-89.

Damaševičius, R. et al., 2020. LITNET-2020: An Annotated Real-World Network Flow Dataset for Network Intrusion Detection. *Electronics,* May, Volume 9, p. 800.

Devara, G., n.d. *tutorialspoint.* [Online]
Available at: https://www.tutorialspoint.com/why-do-we-use-pandas-in-python
[Accessed 31 1 2023].

Duggal, N., 2022. *simplilearn.* [Online]
Available at: https://www.simplilearn.com/advantages-of-decision-tree-article
[Accessed 31 1 2023].

Gairola, S., n.d. *ksolves.* [Online]
Available at: https://www.ksolves.com/blog/big-data/spark/apache-spark-benefits-reasons-why-enterprises-are-moving-to-this-data-engineering-tool
[Accessed 31 1 2023].

Gillis, A. S., n.d. *techtarget.* [Online]
Available at: https://www.techtarget.com/searchenterpriseai/definition/data-splitting
[Accessed 31 1 2023].

Goyal, C., 2021. *analyticsvidhya.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/04/sklearn-objects-fit-vs-transform-vs-fit_transform-vs-predict-in-scikit-learn/#:~:text=%E2%80%93%20It%20is%20used%20on%20the,to%20scale%20our%20test%20data.
[Accessed 31 1 2023].

Jongsuebsuk, P., Wattanapongsakorn, N. & Charnsripinyo, C., 2013. *Real-time intrusion detection with fuzzy genetic algorithm.* s.l., s.n., pp. 1-6.

Lakshmanarao, A., Srisaila, A. & Ravi Kiran, T. S., 2022. *Machine Learning and Deep Learning framework with Feature Selection for Intrusion Detection.* s.l., s.n., pp. 1-5.

Malek, Z. S., Trivedi, B. & Shah, A., 2020. *User behavior Pattern -Signature based Intrusion Detection.* s.l., s.n., pp. 549-552.

Mishra, A., 2018. *towardsdatascience.* [Online]
Available at: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
[Accessed 31 1 2023].

Mohamed, E., 2022. *kaggle.* [Online]
Available at: https://www.kaggle.com/code/essammohamed4320/intrusion-detection-system-with-ml-dl/notebook
[Accessed 14 12 2022].

Pérez-Díaz, J. A., Valdovinos, I. A., Choo, K.-K. R. & Zhu, D., 2020. A Flexible SDN-Based Architecture for Identifying and Mitigating Low-Rate DDoS Attacks Using Machine Learning. *IEEE Access,* Volume 8, pp. 155859-155872.

Priya, S. S., Sivaram, M., Yuvaraj, D. & Jayanthiladevi, A., 2020. *Machine Learning based DDOS Detection.* s.l., s.n., pp. 234-237.

Saxe, J. & Berlin, K., 2015. *Deep neural network based malware detection using two dimensional binary program features.* s.l., s.n., pp. 11-20.

Singh, H., n.d. *einfochips.* [Online]
Available at: https://www.einfochips.com/blog/everything-you-need-to-know-about-hardware-requirements-for-machine-learning/
[Accessed 31 1 2023].

Zhang, J., Liu, P., He, J. & Zhang, Y., 2016. *A Hadoop Based Analysis and Detection Model for IP Spoofing Typed DDoS Attack.* s.l., s.n., pp. 1976-1983.

Zhao, S., Chandrashekar, M., Lee, Y. & Medhi, D., 2015. *Real-time network anomaly detection system using machine learning.* s.l., s.n., pp. 267-270.