National College of Ireland

Effectively improving the efficiency and performance of an intrusion detection system using hybrid machine learning models

MSc Internship
MSc in Cyber Security

Sumanth Kumar Alladi
Student ID: X18108377

School of Computing
National College of Ireland

Supervisor: Michael Pantridge

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sumanth Kumar Alladi
Student ID:	X18108377
Programme:	MSc in Cyber Security **Year:** 2020
Module:	Academic Internship
Supervisor:	Michael Pantridge
Submission Due Date:	17/08/20
Project Title:	Effectively improving the efficiency and performance of an intrusion detection system using hybrid machine learning models
Word Count:	4751 **Page Count:** 15

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.
I agree to an electronic copy of my thesis being made publicly available on NORMA the National College of Ireland's Institutional Repository for consultation.

Signature: Sumanth Alladi

Date: 28/09/2020

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):	

Effectively improving the efficiency and performance of an intrusion detection system using hybrid machine learning models

Sumanth Alladi
X18108377

Abstract

Due to widespread usage of internet we need a proper security network which plays a crucial role by securing the information, as the usage of internet increasing the attackers are also widely increasing. Though we have lot of security systems, hackers are actively using new techniques to overcome this security. Intrusion Detection System (IDS) is a system that provides a security layer to the organization network and it plays a crucial role by blocking the malicious attacks at the initial point of the organization. Here in this research I proposed an IDS hybrid model with Logistic regression with K-means clustering and MLP (Multi-Layer Perceptron) with K-Means clustering. I had chosen NSL-KDD dataset to demonstrate the working of algorithm by testing the dataset and to show the difference between malicious and normal flow of network traffic.

Keywords: *Intrusion Detection System, hybrid algorithms, k-means clustering and MLP algorithm.*

1 Introduction

According to Palo Alto Networks, An Intrusion Detection System (IDS) is a system which is developed to provide network security to vulnerability exploits against a target application or computer. IDS will find the threats in addition to IDS, IPS (Intrusion Prevention System) has an ability to block the threats and become the powerful deployment option for IDS/IPS technologies. The main purpose of the IDS is to monitor the flow of traffic by TAP or SPAN port to analyze a copy of inline traffic stream in this way IDS never impact the traffic flow performance on the network. Because IDS will monitor the flow of network traffic it is also called as listen-only device, IDS will monitor the traffic and reports to the administrator if there is any flaws, it don't have right to take its own decision only admin have right to do once the report made by IDS.
Figure 1. explains the network diagram of an organization.

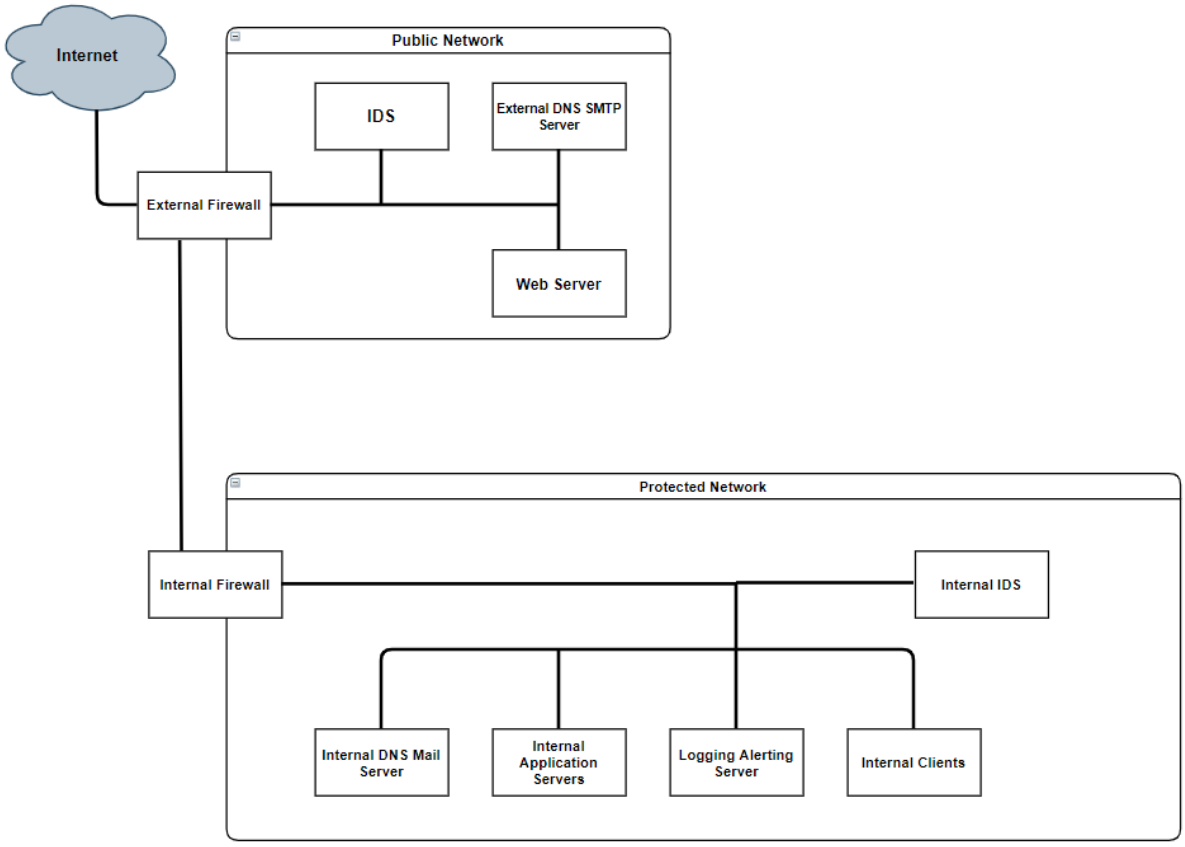

Figure 1: General Network Diagram of an organization

According to SearchSecurity, Intrusion Detection System (IDS) are used to detect and report the anomalies before the hackers really damage the network and the main aim is also to apprehend the hackers. IDS can be 2 types

1. Host based
2. Network based

Host based Intrusion Detection Systems (HIDSs)

From Sciencedirect.com it is believed that these are the applications that operates on the computers where the data of those computers are already collected. The analysis of this HIDSs are in very high level in detailed as the data is already collected so the level of work done by HIDSs are accurate and high. HIDSs are privy to the outcome of the attack attempted by hacker because they already have direct access and monitor to the targeted system.

Host based intrusion detection system uses a unique technique of monitoring the traffic, which is called as TRIPWIRE. HIDSs can use the information sours of OS audits trail which is generated within the kernel (inner level of operating system) so it's more detailed and better protected.

Network Based Intrusion Detection Systems (NIDSs)

Network Based Intrusion Detection System is completely deferent from Host based. Here in NIDSs the main chore is to scan the network packets at router level or host-level. Based on

this scan if any malicious or anomalies found then NIDS will trigger a false alarm or makes administrator alert by passing the message.

In present days due to high growth in internet usage, this network-based IDS are quite common and getting popular. Here in this paper we are also using NIDSs.

Network based intrusion detection system further has 2 types they are

1. Signature-based intrusion detection system (SIDS)
2. Anomaly-based intrusion detection system (AIDS)

Signature based intrusion system has some pre-defined instructions like what kind of traffic is allowed into the network, based of the given instruction commands this SIDS will monitor the traffic and allows into network only if the traffic reaches the required criteria(as per given commands). These are more likely Anti-virus software's.

Anomaly based intrusion detection system monitors the system traffic by comparing the traffic with baseline of the system like suitable bandwidth, protocol, ports, and other devices. This type of IDS mostly uses some machine learning to establish a baseline and accompanying security policies, based on this policy it alerts IT team if any suspicious activity or policy violation appeared.

In the proposed paper, I used logistic regression, MLP algorithm and K-means clustering algorithms. Logistic regression is a machine learning algorithm which is used to assign observations to a discreate set of classes, it is a predictive analysis algorithm based on the concept of probability. The cost function of this algorithm is defined as "Sigmoid Function" and the hypothesis of logistic regression trends to limit of cost function between 0 and 1. Sigmoid function is the function that maps any real value within the dataset to another value between 0 and 1. The reason I chosen this algorithm is because of high volume datasets can be easily analyzed using this logistic regression algorithm.

Other algorithms which I used was MLP multilayer perceptron it is a feedforward artificial neural network (ANN) which is a supervised algorithm which uses a learning technique called backpropagation for training. K-Means clustering algorithm is one of the simplest unsupervised machine learning algorithms, these algorithms makes inferences from dataset using input vectors without referring to known outcomes. By using above 3 algorithms I proposed a hybrid algorithm to increase the accuracy efficiency.

In this research work I have used NSL-KDD dataset which is best suitable for IDS research work. This dataset is split into 2 parts for training and testing purposes because all supervised algorithms must be trained by using training dataset then when the testing dataset fed to the algorithms, they give optimal results

2 Related Work

In this section, we are going to observe the previous studies done by the researchers on Intrusion Detection System (IDS) using different types of machine learning algorithms in multiple situations. By doing this we can perceive what they were able to discover and what were the challenges faced by them. In the first section, the research on types of IDS will be discussed, where the behaviour of IDS is observed in different types of situations based on

the organization. Next section will be about the behaviour of the machine learning algorithms in the IDS systems. Features and data extraction have been observed in the next section where the researchers followed the procedures to extract raw data from the network traffic and convert it into meaningful data points. After that, feature selection has been discussed with its advantages and disadvantages. And in the final section, the observations on the NSL-KDD dataset has been done to get the in-depth information.

2.1 Intrusion Detection System (IDS)

According to the article by Sarmah, A (2020), e-Business has been flourishing due to the exponential growth of the internet. E-Business can take advantage with more and more people connected to the internet. According to him, there are two sides to that business model. The good side is that it brings a lot of opportunities to the organization from the internet, but the bad side is that it also brings a lot of risks with many bad actors present on the internet. The information provided by the organizations is easily available to both harmless and harmful internet users. Hence, the importance of having an IDS systems on the organization's network is a must where it not only monitors and analyzes the network traffic and systems for the possible attacks originating from the outside, but also any misuse or attacks originating from the inside the organization. Another article by Ashoor, A (2011), the behaviour of different types of IDS systems when large data is introduced to them. How the IDS system filters out the noise in the traffic data and applies feature selection techniques to solve the issue of which data points are genuine and which of them are white noise.

There are two ways to tackle the issue of increasing the detection rate in the IDS systems they are by using machine learning algorithm models or by using statistical analysis methods. By the observations, it has been deduced that statistical analysis method takes a lot of time and effort to detect any anomalies in the network traffic, which can lead up to a detrimental effect for the organization. Whereas, using machine learning algorithms has proven much more efficient and quicker for the same task. It has been observed that the prediction and decision-making qualities increases when a large amount of data gets introduced to the system.

2.2 Machine Learning Algorithms

In the research paper written by Tahir, H & Said, A (2016), the Anomaly Based Intrusion Detection (ABID) was taken and observed. Here, they were able to improve the performance of the IDS by creating the hybrid model of K-Means Clustering and Naïve Bayes Classifier. They were able to achieve the detection rate of 99.3% and an accuracy of 99.5% against ISCX 2012 dataset. It also showed that the false alarm rate was reduced significantly which gave optimal results. Waikato Environment for Knowledge Analysis (WEKA) data mining tool was used as it supports many machine learning algorithms and data mining tasks such as clustering, preprocessing, regression, feature selection and many more to name a few.

In another research paper by Eslamnezhad, M (2014), a comparison between the IDS systems was observed where one IDS has only K-Means clustering algorithm in it and the other one has MinMax K-Means clustering algorithm. NSL-KDD dataset was used to introduce the network traffic data points. Observing the first IDS with only K-Means, it

resulted in a lacklustre output where the sensitivity and detection rate was average with higher false positive detection rates. It was also observed that K-Means had three main drawbacks which are force assignment problem, no class problem and class dominance problem. But when they passed the same dataset through second IDS which had the MinMax K-Means implementation, they were able to observe that the sensitivity and detection rates of the IDS were much higher with low false positive detection rates. This gave me the overall idea of how K-Means clustering algorithm works and what can be done to make more improvements on it.

Alzahrani, A (2019) made the comparison between Structural Sparse Logistic Regression (SSPLR) model and Support Vector Machine (SVM) model in the IDS system to see which one will give the optimal result when the NSL-KDD dataset was passed through them. It was observed that SSPLR performed better than SVM model when it came to feature selection technique. This was mainly due to SSPLR being one stage method whereas SVM is a two-stage method. Due to this SVM required much more training and testing time with much more requirement of the computational power. With Sparse technique applied by SSPLR, coefficients provided by help IDS to determine how a group of features affect the probability of specific security attack classes to occur. The only challenge that was observed is the labelling of the real-time network traffic will be very time consuming and costly.

Another research paper that helped me to get the idea of using the MLP model was by Yao, Y (2006), where they propose a hybrid model between MLP and Chaotic Neural Network. Comparing with the MLP only model, the hybrid model was able to achieve an improvised detection of the time-delayed attacks efficiently with lower false alarm rate when detecting novel attacks. Here they used DARPA 1998 dataset to pass through both the models.

2.3 Feature Extraction and Data Collection

Here in this section, the raw data gets collected from the network and then gets changed to a meaningful and easily recognizable data. This is done by using feature extraction techniques. Researchers either created a network environment or use an existing dataset which contains previously recorded network traffic. Pcap files have been created by using tools such as Wireshark or Libcap and then it gets passed through the algorithm model. Softwares such as Netmate is used to convert the captured pcap files to the flow statistics. 45 features such as destination IP, source IP, destination port, source port, average bits per second, the length of the first packet to name a few.

Chen, R (2017) collected the data from the network environment that was created at the university. This data is further converted into a dataset. He chose headers depending on five factors that are destination IP, source IP, destination port, source port and lastly protocol. If the source port or destination port and the protocol were similar, then it was considered as same flow. The researched admitted that the datapoints that were captured were not volatile enough to be considered for other situations. Hence, the recommendation for using a larger dataset such as CTU-13 was made.

Hung, C (2018) compared the datasets of the three Botnet families of Zeus, Virut and Waledac with the pcap file that was created by him to tackle the generalization problem. This pcap file was created in the privately controlled network environment created at the university. A huge amount of traffic was simulated to create a huge pcap file with the size of 120GB. The researched admits that he was not able to replicate real-world network traffic, where there are many types of Botnets. Netmate was used to capture the network traffic and convert it into a pcap file.

2.4 Feature Selection

To get the initial idea of the feature selection method and how it's done, I investigated the research done by Javadpour, A (2017) where KDD99 dataset was used. Two types of feature selection techniques were applied which are linear correlation and mutual information. Out of 41 features present in the dataset, 21 of them were selected and were passed through different classifier algorithms such as Random Forest, CART algorithm, decision tree and neural networks. Neural networks recorded the highest accuracy of 99.98% when compared with others. It was also deduced that INTERACT method played an important part where the importance of a feature was decided depending on its relationship with other features in the dataset.

Study on Principal Component Analysis (PCA) was done by Lakhina, S (2010), where NSL-KDD dataset was taken and was passed through the model to apply PCA. This resulted in the feature reduction which was done by first converting the datapoints into 1s or 0s, which made it simpler to deal with. The goal of PCA is to keep the integrity of the dataset intact without the loss of the original data. Originally the dataset contained 41 features which were then reduced to a total of five classes namely Normal, DoS, Probe, U2L and R2L. By applying this technique, the researcher was able to increase the efficiency and processing time of the machine learning algorithms.

2.5 Dataset

Research on multiple datasets was made by Ring, M (2019), where they found that NSL-KDD dataset was the refined version of the KDDCup 99 dataset. Many researchers who used KDDCup dataset noted that there were many big flaws in it, which in-turn invalidated their outputs. The main anomaly in that dataset was that all the malicious packets have TTL port numbers 126 or 253 assigned to them. On the other hand, all the benign packets were assigned the port numbers either 127 or 254. Which showed a bias towards a particular traffic flow. To tackle this redundant record were removed which resulted in the unbiased results. Plus, the number of train and test data were increased to get more optimal results. This led me to choose this dataset.

3 Research Methodology

In this research work, I have proposed Hybrid models which is used for intrusion detection system (IDS). For the Hybrid model, Logistic Regression with k-means clustering algorithms have been used and for the second hybrid model MLP with k-means clustering algorithms have been used. According to the previous researchers work, it is believed that hybrid model performs better than the single algorithm alone. So, in this research work hybrid models have been implemented against a pre-existing dataset to obtain maximum accuracies. The combination of these algorithms to form hybrid models has never been used before.

To develop Intrusion detection system (IDS), selection of dataset plays crucial role. The selected dataset consists of malicious flow, normal flow, and background flow. To choose a dataset, researcher either need to generate structured dataset in a controlled environment or can select a pre-existing dataset. In this research work, a pre-existing dataset (NSL-KDD) has been used because the process of generating a new dataset tends to get tedious and time consuming. This dataset consists of malicious flow and normal flow. The information regarding selected dataset, per-processing, feature selection, Hybrid model prediction process and the results have been presented below.

In this research work, the way these hybrid models are implemented has been shown in figure 2.

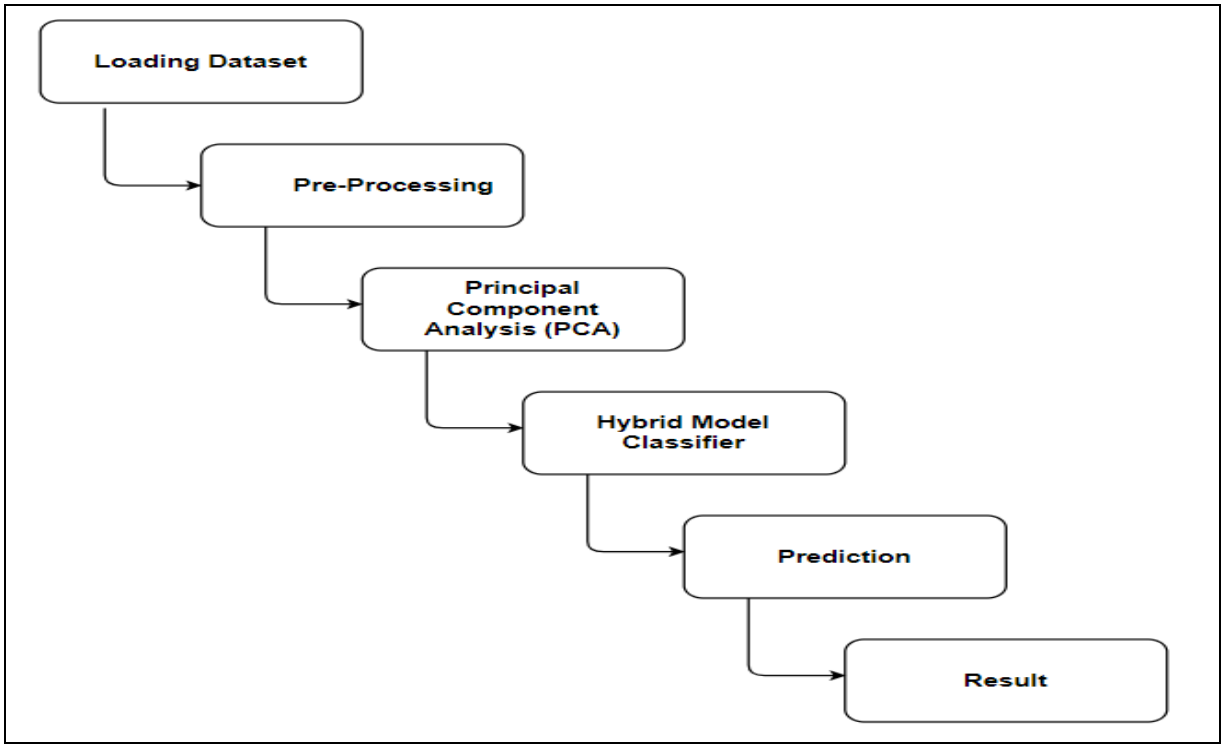

Figure 2: Flow diagram of Hybrid Model

3.1 Dataset Selection

In this research work, I have used a pre-existing dataset which is NSL-KDD. This dataset is best known IDS dataset. Many previous researchers have been used this dataset for their intrusion detection system projects because this dataset contains majority of malicious flow along with normal traffic and background traffic. This dataset is split into two parts, one

part is used for training supervised algorithms and the other part is used for testing the k-means algorithm and already trained supervised algorithm. Around 70% of the dataset has been used for training purpose because the more the supervised algorithm is trained the better test results will be obtained. 30% of dataset is used for testing purpose.

3.2 Pre-processing

In the pre-processing stage, firstly, the dataset will be loaded and split into train dataset and test dataset. By using label encoder existing nominal variable are transformed into integer variables, all these variables are scaling in 0 and 1. 0 defines normal flow and 1 defines malicious flow.

3.3 Feature Selection

After loading the dataset, feature selection comes into the picture. In this research work, principal component analysis (PCA) technique is used for feature selection. By implementing PCA all the unwanted features or labels will be removed. This way the performance of algorithms will be improved.

3.4 Prediction Process

Firstly, the training dataset is given to the supervised algorithms because these algorithms must be trained so that while predicting the test dataset the results will be optimal. Secondly, after training test dataset is fed to the algorithms and the results will be obtained in integer variables. The hybrid model is prepared in such a way that the obtained results will be in 0's and 1's. 0 defines normal flow and 1 defines malicious flow. For IDS, the hybrid model gives accurate results than the actual algorithm.

3.5 Result

After obtaining the results, the performance of all the algorithms used in this research work along with hybrid models have been tabulated and compared.
The result shows Accuracy, Sensitivity and Specificity of all the algorithms and hybrid models.

4 Design Specification

Firstly, Dataset is loaded then in the pre-processing section, the dataset is split into train dataset and test dataset. For training 70% of dataset has been used because more the training for supervised algorithms the better predictions will be obtained when the algorithms run with test dataset. The feature selection is done by using principal component analysis (PCA), it removes all the redundant features or labels.

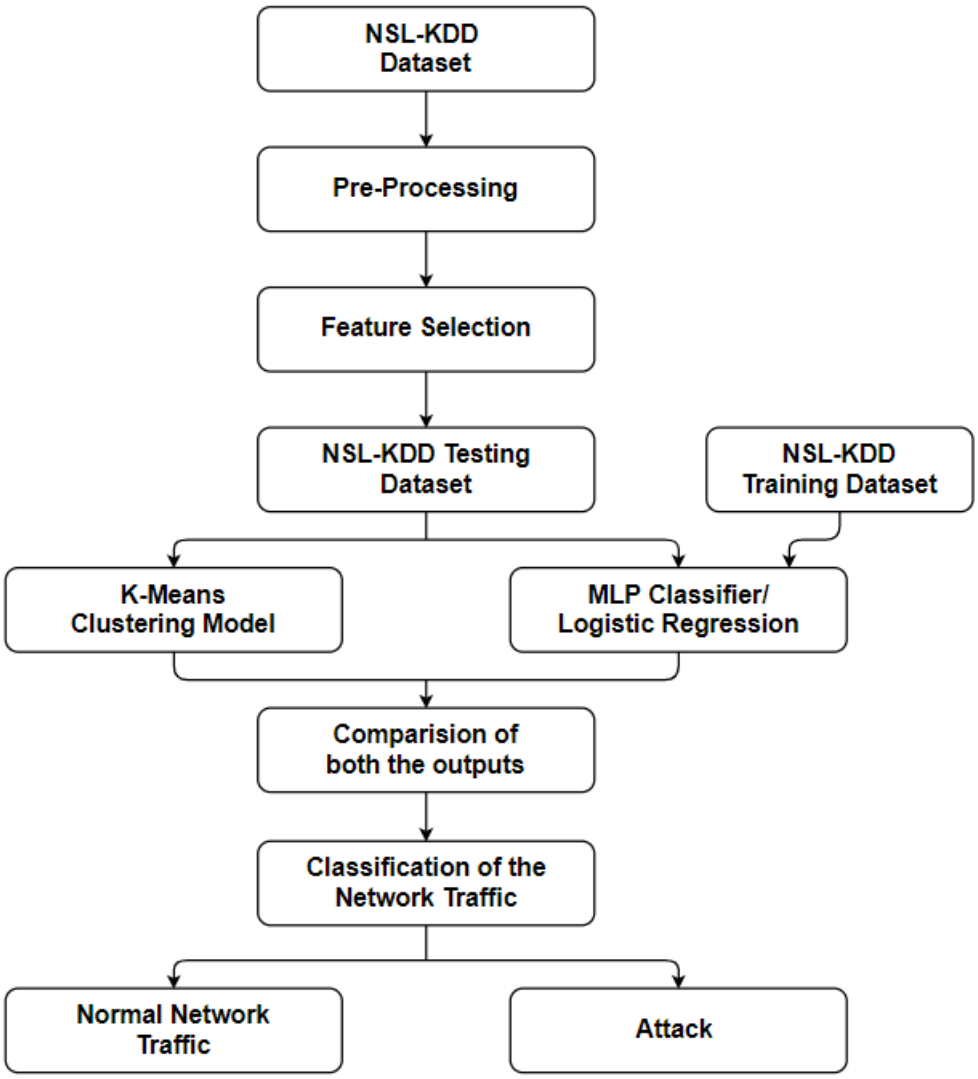

Figure 3: Frame Work

The MLP algorithm and logistic regression algorithms are trained by training dataset. After the training testing dataset is fed to both MLP and LR algorithms. Simultaneously, the test dataset is fed to the K-Means clustering algorithms because the un-supervised algorithms does not need any training, it will find its own path when reading the dataset and predicts accurate results.

In this research work, I have used two hybrid models, one is LR & K-Means and the second is MLP & K-Means. The outputs are compared in such a way that the predicted output from LR algorithm and from K-Means clustering algorithm will be in 0's and 1's. So, if LR algorithm predicts the flow is malicious traffic (1) and K-Means also predicts the flow is malicious traffic (1) then we consider malicious traffic as the final prediction for hybrid model.

"0" means normal flow and "1" means malicious flow.

5 Implementation

5.1 Loading of Dataset

In this research work Scikit-learn machine learning library has been used and it features various classification, regression, and clustering algorithms. Firstly, NSL-KDD dataset is loaded. The dataset is divided into train dataset and test dataset, for training 70% of dataset has been used and for testing 30% of the dataset has been used.

5.2 Feature selection

Feature selection plays an important role when implementing a machine learning algorithm, because by using feature selection we can be able to drop unwanted or unnecessary columns or features.

In this research work, principal component analysis has been used for feature selection. Basically, PCA is a dimensionality reduction method that is used to change large datasets into small datasets. According to Built In, The main purpose of this process is to analyze data faster and easier for machine learning algorithms. By trading little accuracy, we are making dataset simpler to analyze.

After eliminating unwanted features, the dataset is split into train dataset and test dataset. This entire process is carried out in preprocess.py python file. Thereby transforming nominal variables into integer variable by using label encoder. So, the obtained outputs from algorithms varies in 0's and 1's. "0" defines normal traffic and "1" defines malicious traffic.

5.3 Hybridization

The hybridization is carried out in lr_kmeans.py and mlp_kmeans.py python scripts. After training the logistic regression algorithm using train dataset, the test dataset is fed to the logistic regression algorithm and starts predicting malicious flow (1) and normal flow (0). Simultaneously, test dataset is running against K-means clustering algorithm obtaining results in malicious flow (1) and normal flow (0). The hybridization works in such a way that when both the algorithms predict that the flow is malicious (1) then we consider the final result as malicious flow whereas in all other cases we consider the flow is normal or background flow. Similarly, the hybridization is done for MLP and K-Means clustering algorithm.

5.4 Output

Each machine learning python scripts in this research work returns the values of accuracy, specificity, and sensitivity. By using main.py python script, I have tabulated obtained outputs of hybrid models and individual algorithms. Thereby, using a graph the results are compared.

6 Evaluation

In this research work, evaluation is done by comparing obtained outputs such as accuracy, specificity, and sensitivity. By using a confusion matrix these results were obtained. A confusion matrix is a table used to describe the classification algorithms performances and therefore it visualizes the performance of the algorithms.

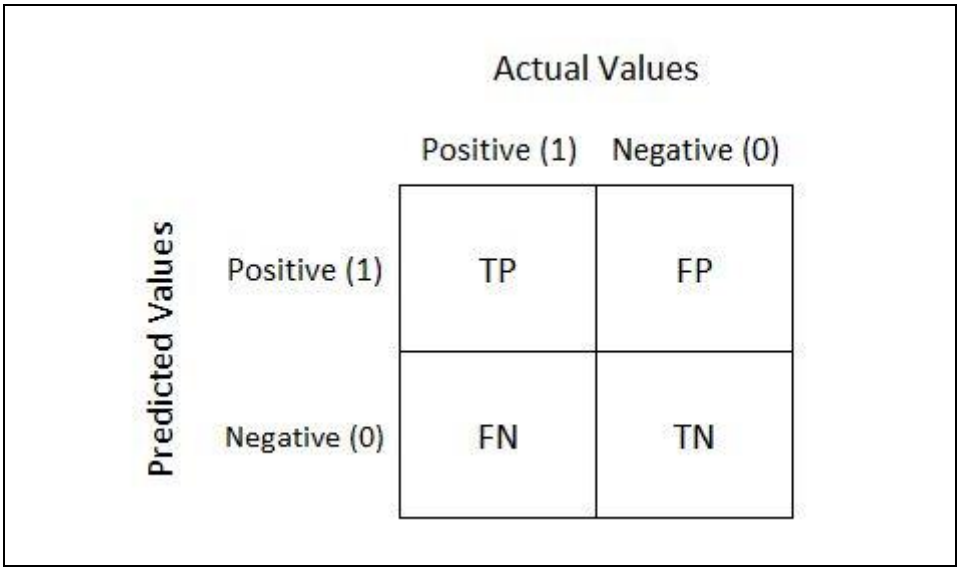

Figure 4: Confusion Matrix

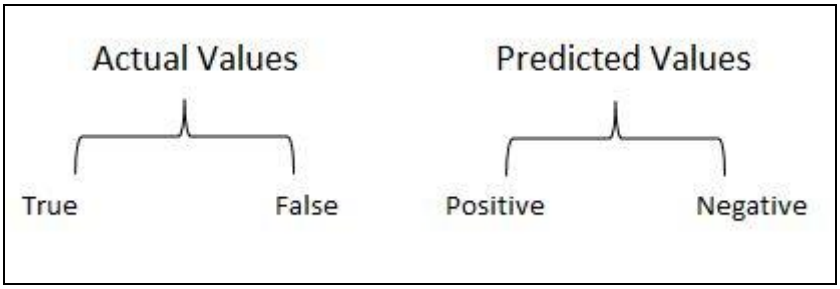

Figure 5: Actual and Predicted values explanation

According to the figures 4 and 5 above, true positive and true negative are when your prediction and actual values are same.

Interpretation:

True Positive – you predict positive and it is true.

True Negative – you predict negative and it is true.

False Positive and False Negative are type 1 and type 2 errors.

6.1 Obtained output

In this research paper, accuracy, sensitivity, and specificity were calculated using confusion matrix.

The formula to calculate accuracy is Accuracy = (TP+TN)/(TP+TN+FP+FN)

Sensitivity is true positive rate and the formula to calculate is **sens** = TP / (TP + FN)

Specificity is true negative rate and the formula to calculate is **spec** = TN / (TN + FP)

Method	Accuracy	Sensitivity	Specificity
MLP	0.848334	0.969825	0.756409
LR	0.846516	0.927394	0.785319
Kmeans	0.724083	0.975798	0.533624
MLP & Kmeans	0.730959	0.565088	0.856464
LR & KMeans	0.847669	0.924305	0.789683

Figure 6: Output Table

According to the above table, the logistic regression and K-Means clustering hybrid model performance is better and improved in all aspects. Whereas the MLP and K-Means hybrid model performance is slightly dropped. MLP and k-means hybrid model gives high specificity values compared to all other algorithms used in this research work.

6.2 Performance Graph

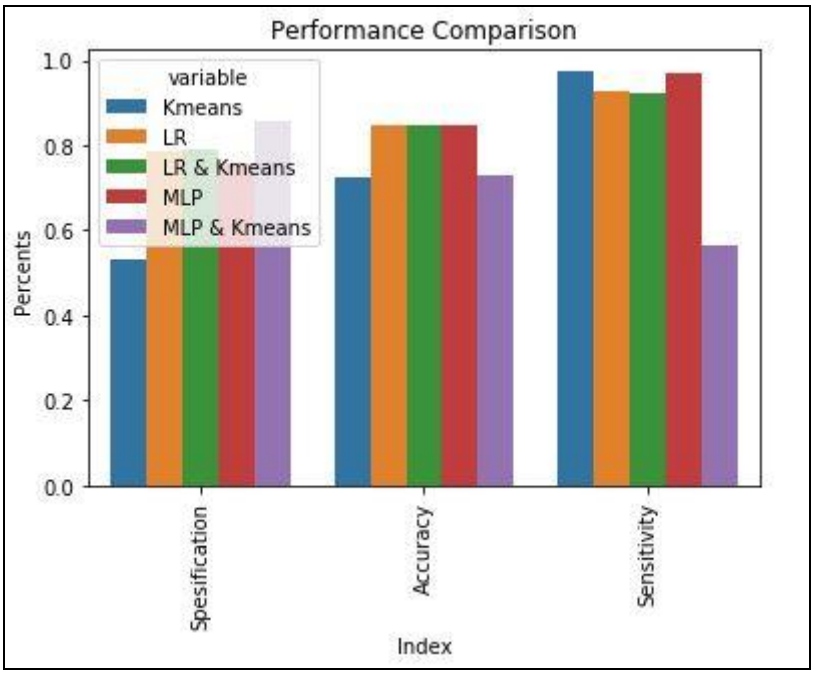

Figure 7: Performance Comparision

The above Figure 7. shows the performance comparison of all three algorithms along with two hybrid models. The hybrid model's performance is better when compared with individual algorithm alone.

6.3 Findings

- The performance of LR & K-means hybrid model is better when compared with other hybrid model and individual algorithms.
- Specificity values are high for both the hybrid models.
- The performance of MLP &K-means hybrid model is slightly poor when compared with LR & K-means hybrid model.

6.4 Discussion

In this research work it has been proved that the hybrid model's performance is better when compared with single algorithm alone. By implementing principal component analysis, it is proved that the overall performance has been increased for each and every algorithm. The proposed logistic regression and K-Means clustering hybrid model giving high performance specs such as accuracy 84.7%, Sensitivity 92.4%, and finally specificity 78.9%. In this research work two hybrid models were proposed, the logistic regression with k-means clustering hybrid model giving high accuracy and sensitivity than the MLP with k-means clustering hybrid model. Whereas MLP & k-means model is giving higher specificity rate than all the individual algorithms and hybrid model implemented in this research work.

Hybrid Model	Accuracy	Sensitivity	Specificity
MLP & k-means	73.09%	56.5%	85.6%
LR & k-means	84.7%	92.4%	78.9%

Table: Outputs of Both Hybrid Models

The above table shows the outputs obtained by two hybrid models. The LR & k-means hybrid model giving optimal results.

7 Conclusion and Future Work

This research work is aimed at increasing the performance of intrusion detection system (IDS) by implementing hybrid machine learning model. A substantial understanding of domain is obtained by carefully going through previous researchers work in this field. The proposed hybrid model LR & K-Means algorithm is performing better than other algorithms and MLP & K-Means hybrid model. Nonetheless the second hybrid model, MLP &K-Means hybrid model performance is high in calculating specificity.

In this research work, for the hybrid model one supervised and one un-supervised algorithms were used. For future work, a hybrid model can be made using three algorithms for example two supervised and one unsupervised or two unsupervised and one supervised algorithm. Supervised algorithms like boosting and support vector machines and unsupervised algorithm like K-Means clustering can be used for future work.

References

"What is an Intrusion Detection System?", Palo Alto Networks, 2020. [Online]. Available: https://www.paloaltonetworks.com/cyberpedia/what-is-an-intrusion-detection-system-ids#:~:text=An%20Intrusion%20Detection%20System%20(IDS,a%20target%20application%20or%20computer. [Accessed: 17- Aug- 2020].

"What is an Intrusion Detection System (IDS) and How Does it Work?", SearchSecurity, 2020. [Online]. Available: https://searchsecurity.techtarget.com/definition/intrusion-detection-system. [Accessed: 17- Aug- 2020].

"Host-Based Intrusion Detection Systems - an overview | ScienceDirect Topics", Sciencedirect.com, 2020. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/host-based-intrusion-detection-systems. [Accessed: 17- Aug- 2020].

"Network-based IDS", Web.mit.edu, 2020. [Online]. Available: https://web.mit.edu/rhel-doc/4/RH-DOCS/rhel-sg-en-4/s1-ids-net.html. [Accessed: 17- Aug- 2020].

A. Sarmah, "Intrusion Detection Systems: Definition, Need and Challenges", 2020. [Accessed 16 August 2020]

A. Ashoor and S. Gore, "Importance of Intrusion Detection System (IDS)", 2011. [Accessed 16 August 2020].

H. Tahir and A. Said, "Improving K-Means Clustering Using Discretization Technique in Network Intrusion Detection System", 2016. [Accessed 16 August 2020].

M. Eslamnezhad and A. Varjani, "Intrusion Detection Based on MinMax K-means Clustering", 2014. [Accessed 16 August 2020].

A. Alzahrani and R. Shah, "A novel method for feature learning and network intrusion classification", 2019. [Accessed 16 August 2020].

Y. Yao, Y. Wei, F. Gao and Y. Ge, "Anomaly Intrusion Detection Approach Using Hybrid MLP/CNN Neural Network," Sixth International Conference on Intelligent Systems Design and Applications, Jinan, 2006, pp. 1095-1102, doi: 10.1109/ISDA.2006.253765. [Accessed 16 August 2020].

R. Chen, W. Niu, X. Zhang, Z. Zhuo and F. Lv, "An Effective Conversation-Based Botnet Detection Method", Mathematical Problems in Engineering, vol. 2017, pp. 1-9, 2017. Available: 10.1155/2017/4934082 [Accessed 16 August 2020].

C. Hung and H. Sun, "A Botnet Detection System Based on Machine-Learning using Flow-Based Features", 2018. [Accessed 16 August 2020].

A. Javadpour, S. Kazemi Abharian and G. Wang, "Feature Selection and Intrusion Detection in Cloud Environment Based on Machine Learning Algorithms", 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), 2017. Available: 10.1109/ispa/iucc.2017.00215 [Accessed 16 August 2020].

S. Lakhina, S. Joseph and B. Verma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", 2010. [Accessed 16 August 2020].

M. Ring, S. Wunderlich, D. Scheuring, D. Landes and A. Hotho, "A survey of network-based intrusion detection data sets", Computers & Security, vol. 86, pp. 147-167, 2019. Available: 10.1016/j.cose.2019.06.005. [Accessed 16 August 2020]

"A Step by Step Explanation of Principal Component Analysis", Built In, 2020. [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis. [Accessed: 17- Aug- 2020].

"Confusion Matrix in Machine Learning - GeeksforGeeks", GeeksforGeeks, 2020. [Online]. Available: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/. [Accessed: 17- Aug- 2020].