IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF CYBER ATTACKS USING MACHINE LEARNING TECHNIQUE

Mr.S.S.Vasantha Raja [1],Aakash B [2,]Avinash M[3],Gokul S[4]

[1]Asst.Prof.,[2,3,4]Student

Department of Computer Science and Engineering

PERI INSTITUTE OF TECHNOLOGY

Abstract

Cyber-attack, via cyberspace, targets an enterprise's use of cyberspace for the purpose of disrupting, disabling, destroying, or maliciously controlling a computing environment/infrastructure; or destroying the integrity of the data or stealing controlled information. The state of the cyberspace portends uncertainty for the future Internet and its accelerated number of users. New paradigms add more concerns with big data collected through device sensors divulging large amounts of information, which can be used for targeted attacks. Though a plethora of extant approaches, models and algorithms have provided the basis for cyber-attack predictions, there is the need to consider new models and algorithms, which are based on data representations other than task-specific techniques. However, its non-linear information processing architecture can be adapted towards learning the different data representations of network traffic to classify type of network attack. In this, we are modeling cyber-attack prediction as a classification problem, Networking sectors have to predict the type of Network attack from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information’s like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the type cyber Attacks. We classify four types of attacks are DOS Attack, R2L Attack, U2R Attack, Probe attack. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with entropy calculation, precision, Recall, F1 Score, Sensitivity, Specificity and Entropy.

Introduction

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms.

It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points.

Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

The approach taken by most organizations to counter cyber attacks is defensive and reactionary. Threats are only removed and analyzed once they are detected; at which point, the harm is done—the network has already been breached and valuable information compromised. Intrusion detection and prevention, such as anti-virus software and firewalls combined with access controls such as passwords, are the common tools and measures employed by the majority of organizations.

Considering how sophisticated and multi-faceted cyber crimes have been in recent years, and how numerous the attacks that don’t make the tabloids are, however, it’s safe to say reactionary responses are damage control measures, at best, and are largely ineffective.The future of cyber security isn’t as bleak as we’ve probably made it sound, however. With the rapid development of AI, ML and quantum encryption, many new ways to combat cyber attacks are emerging. Cyber security has a big role to play in the future of our industries and the government. With this understanding, many of our greatest minds are tackling this problem and have already made promising progress.

Literature Survey

The process of prediction analysis is a process of using some method or technology to explore or stimulate some unknown, undiscovered or complicated intermediate processes based on previous and present states and then speculated the results. In an early warning system, accurate prediction of DoS attacks is the prime aim in the network offence and defense task. Detection based on abnormity is effective to detect DoS attacks. A various studies focused on DoS attacks from different respects. However, these methods required a priori knowledge being a necessity and were difficult to discriminate between normal burst traffics and flux of DoS attacks. Moreover, they also required a large number of history records and cannot make the prediction for such attacks efficiently.

Based on data from flux inspecting and intrusion detection, it proposed a prediction model of DOS attack's distribution discrete probability based on clustering method of genetic algorithm and Bayesian method and the clustering problem first, and then utilizes the genetic algorithm to implement the optimization of clustering methods. Based on the optimized clustering on the sample data, we get various categories of the relation between traffics and attack amounts, and then build up several prediction sub-models about DoS attack. Furthermore, according to the Bayesian method and deduce discrete probability calculation about each sub-model and then get the distribution discrete probability prediction model for DoS attack.

This paper begins with the relation exists between network traffic data and the amount of DoS attack, and then proposes a clustering method based on the genetic optimization algorithm to implement the classification of DoS attack data. This method first gets the proper partition of the relation between the network traffic and the amount of DoS attack based on the optimized clustering and builds the prediction sub-models of DoS attack. Meanwhile, with the Bayesian method, the calculation of the output probability corresponding to each sub-model is deduced and then the distribution of the amount of DoS attack in some range in future is obtained.

Due to emergence of internet on mobile phone, the different social networks such as on social networking sites, blogs, opinion, ratings, review, serial bookmarking, social news, media sharing, Wikipedia led the people to disperse any kind of information very easily. Rigorous analysis of these patterns can reveal some very undisclosed and important information explicitly whether that person is conducting malignant or harmless communications with a particular user and may be a reason for any kind of socio technical attacks.

From the above simulation done on CDR, it may be concluded that if this kind of simulation applied on networks based on the internet and if we are in the position to get the data which could be transformed in transition and emission matrix then several kind of prediction may be drawn which will be helpful to take our decisions.[2]

Intrusion detection systems (IDS) are used to detect the occurrence of malicious activities against IT system. Through monitoring and analyzing of IT system activities the malicious activities will be detected. In ideal case IDS generate alert(s) for each detected malicious activity and store it in IDS database. Some of stored alerts in IDS database are related. Alerts relations are differentiated from duplication relation to same attack scenario relation. Duplication relation means that the two alerts generated as a result of same malicious activity. Where same attack scenario relation means that the two related alert are generated as a result of related malicious activities.

Attack scenario or multi-step attack is a set of related malicious activities run by same attacker to reach specific goal. Normal relation between malicious activities belong to same attack scenario is causal relation. Causal relation means that current malicious activity output is pre-condition to run the next malicious activity. Possible multi-step attack against a network start with information gathering about network and the information gathering is done through network Reconnaissance and fingerprinting process.

Through reconnaissance network configuration and running services are identified. Through fingerprint process Operating system type and version are identified. Propose a real time prediction methodology for predicting most possible attack steps and attack scenarios. Proposed methodology benefits from attacks history against network and from attack graph source data. it comes without considerable computation overload such as checking of attack plans library. It provides parallel prediction for parallel attack scenarios.

Possible third attack step is to identify attack plan based on the modeled attack graph in the past step. The attack plan usually will include the exploiting of a sequence of founded vulnerabilities. Mostly this sequence is distributed over a set of network nodes. This sequence of nodes vulnerabilities is related through causal relation and connectivity. Lastly Attacker start orderly exploits the attack scenario sequences till reaching his/her goal. Attack plan consist of many correlated malicious activities end up with attacking goal.

The prediction results reflect the security situation of the target network in the future, and security administrators can take corresponding measures to enhance network security according to the results. To quantitatively predict the possible attack of the network in the future, attack probability plays a significant role. It can be used to indicate the possibility of invasion by intruders.[4]

Graphical models such as attack graphs become the main-stream approach. Attack graphs which capture the relationships among vulnerabilities and exploits show us all the possible attack paths that an attacker can take to intrude all the targets in the network. The traffics to different hosts or servers may differ from each other. The hosts or servers with big traffic may be more risky since they are often important hosts or servers, and intruders may have more contacts and understanding with them. In our cyber-attacks prediction model, they used attack graph to capture the vulnerabilities in the network.

In addition we consider 3 environment factors that are the major impact factors of the cyber-attacks in the future. They are the value of assets in the network, the usage condition of the network and the attack history of the network. Cyber-attacks prediction is an important part of risk management. Existing cyber-attacks prediction methods did not fully consider the specific environment factors of the target network, which may make the results deviate from the true situation. In this paper, we propose cyber-attacks prediction model based on Bayesian network. We use attack graphs to represent all the vulnerabilities and possible attack paths. Then we capture the using environment factors using Bayesian network model. Cyber-attacks predictions are performed on the constructed Bayesian network.

This paper begins with the relation exists between network traffic data and the amount of DoS attack, and then proposes a clustering method based on the genetic optimization algorithm to implement the classification of DoS attack data.[5] This method first gets the proper partition of the relation between the network traffic and the amount of DoS attack based on the optimized clustering and builds the prediction sub-models of DoS attack.

Meanwhile, with the Bayesian method, the calculation of the output probability corresponding to each sub-model is deduced and then the distribution of the amount of DoS attack in some range in future is obtained. This paper describes the clustering problem first, and then utilizes the genetic algorithm to implement the optimization of clustering methods. Based on the optimized clustering on the sample data, we get various categories of the relation between traffics and attack amounts, and then build up several prediction sub-models about DoS attack. Furthermore, according to the Bayesian method, we deduce discrete probability calculation about each sub-model and then get the distribution discrete probability prediction model for DoS attack.

System Design

System Architecture Diagram

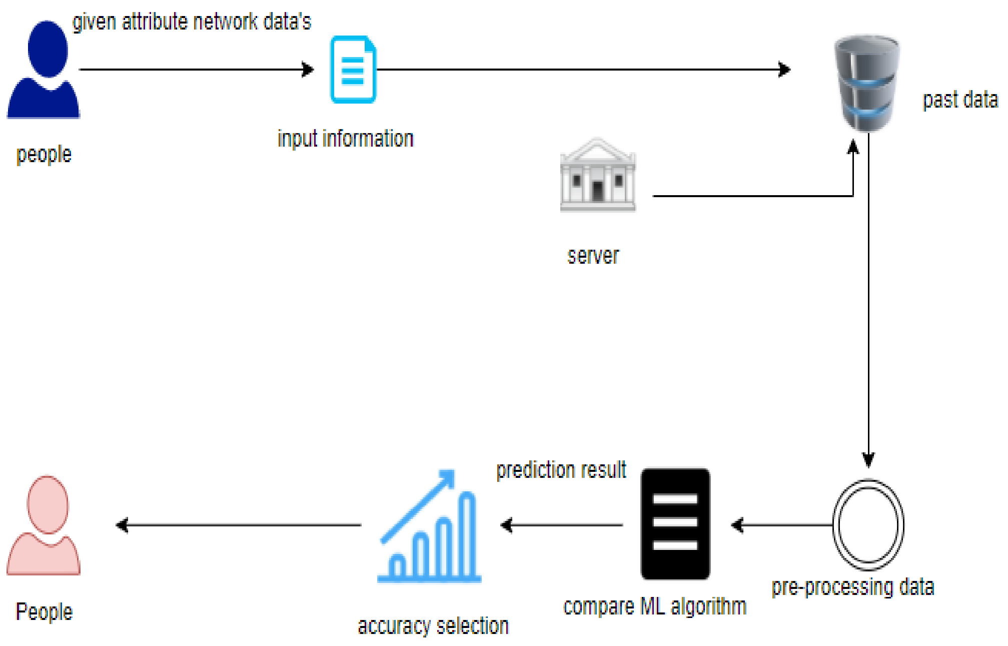


Fig 1: system design

IJCRT

System architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system.

The user will give the input data and those input data will be pre-process using the past dataset and then evaluating those input data using different machine learning algorithms. And then the high accuracy of the algorithm will be given as a GUI output

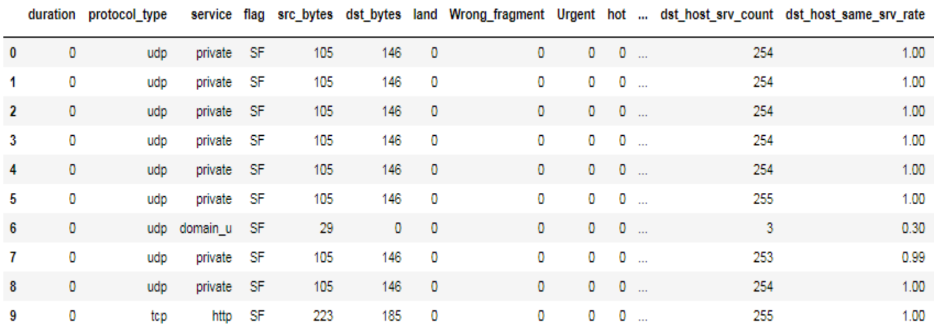
The performance is not good and its get complicated for other networks.The performance metrics like recall F1 score and comparison of machine learning algorithm is not done.

System Implementation

Variable Identification Process / data validation process

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. (For example to show the data type format of given dataset)



	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	Wrong_fragment	Urgent	hot	...	dst_host_srv_count	dst_host_same_srv_rate
0	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
1	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
2	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
3	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
4	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
5	0	udp	private	SF	105	146	0	0	0	0	...	255	1.00
6	0	udp	domain_u	SF	29	0	0	0	0	0	...	3	0.30
7	0	udp	private	SF	105	146	0	0	0	0	...	253	0.99
8	0	udp	private	SF	105	146	0	0	0	0	...	254	1.00
9	0	tcp	http	SF	223	185	0	0	0	0	...	255	1.00

Fig 2: Given Data Frame

Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning models and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Exploration data analysis of visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance

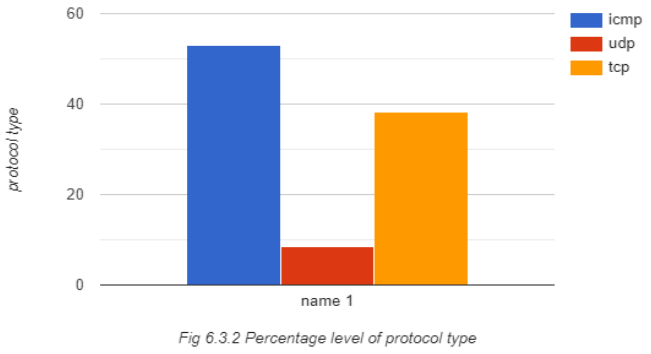


Fig 3: Percentage level of protocol type

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

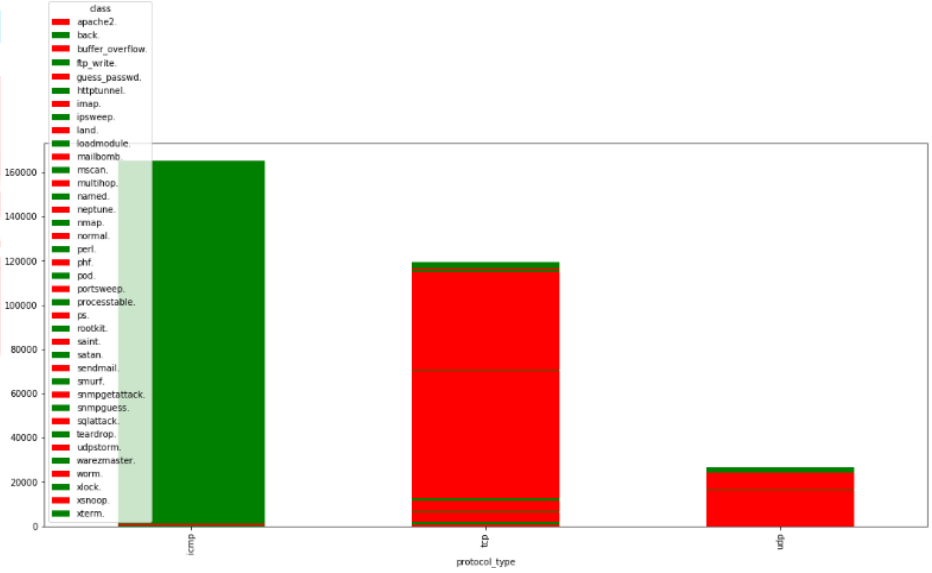


Fig 4: Comparison of service type and protocol type

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results.

Even before predictive models are prepared on training data, outliers can result in misleading representations and in turn misleading interpretations of collected data. Outliers can skew the summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatter plots, compressing the body of the data. Finally, outliers can represent examples of data instances that are relevant to the problem such as anomalies in the case of fraud detection and computer security.

It couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows:

- Reserve some portion of sample data-set.
- Using the rest data-set train the model.
- Test the model using the reserve portion of the data-set.

Advantages of train/test split

- This runs K times faster than Leave One Out cross-validation because K-fold cross-validation repeats the train/test split K-times.
- Simpler to examine the detailed results of the testing process.

Advantages of cross-validation:

- More accurate estimate of out-of-sample accuracy.
- More "efficient" use of data as every observation is used for both training and testing.

Data Pre-processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner.

Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository.

Evaluation Metrics

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

True Positive Rate (TPR) = TP / (TP + FN)

False Positive Rate (FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation: Accuracy = (TP + TN) / (TP + TN + FP + FN) Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. **Precision:** The proportion of positive predictions those are actually correct.

Precision = TP / (TP + FP) Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict) Recall = TP / (TP + FN)

Recall (Sensitivity): Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

General Formula: F-Measure = 2TP / (2TP + FP + FN)

F1-Score Formula: F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Performance Analysis

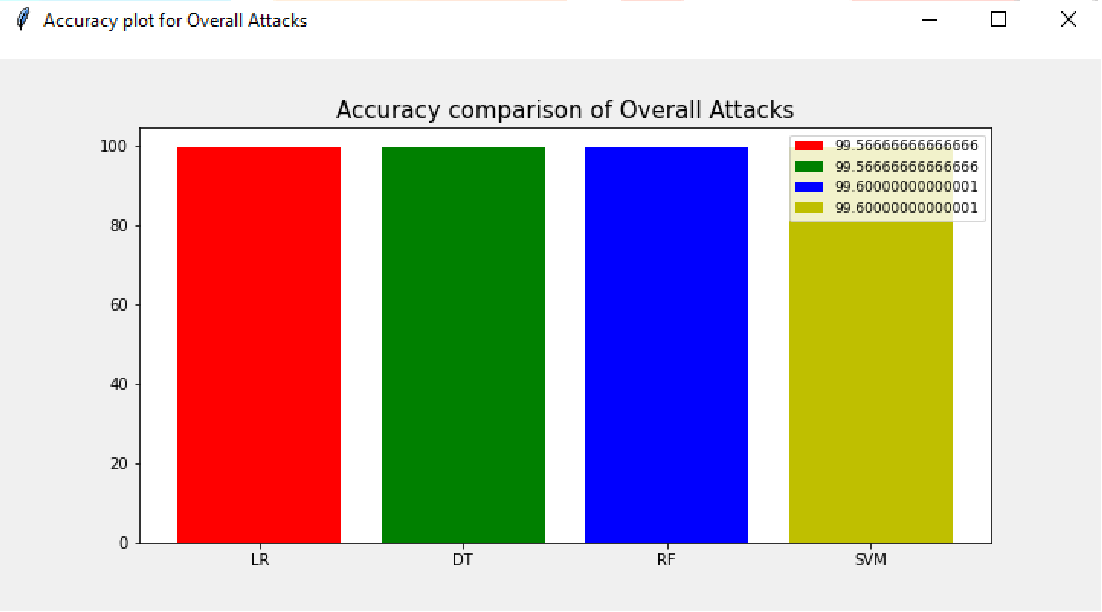


Fig 5: Accuracy Comparison

We have taken 4 types of Algorithms namely Logistic Regression [LR], Decision Tree [DT], Random Forest [RF], Support Vector Machine [SWM]. After continuous evaluation of the overall attacks with the algorithms, it is proved that **Random Forest Algorithm and Support Vector Machine** have the highest percentage in overall comparison.

Conclusion

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be finding out by comparing each algorithm with type of all network attacks for future prediction results by finding best connections. This brings some of the following insights about diagnose the network attack of each new connection. To presented a prediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection. It can be inferred from this model that, area analysis and use of machine learning technique is useful in developing prediction models that can helps to network sectors reduce the long process of diagnosis and eradicate any human error.

Future Enhancements

Network sector want to automate the detecting the attacks of packet transfers from eligibility process (real time) based on the connection detail. To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in Artificial Intelligence environment.

References

1. Anomaly Detection on Attributed Networks via Contrastive Self-Supervised Learning, 2021.
2. A Prediction Model of DoS Attack’s Distribution Discrete Probability, 2008.
3. Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network, 2014.
4. New Attack Scenario Prediction Methodology, 2013.
5. A study on reduced support vector machines, 2003.
6. Cyber Attacks Prediction Model Based on Bayesian Network, 2012.
7. Adversarial Examples: Attacks and Defenses for Deep Learning, 2019.
8. A Prediction Model of DoS Attack’s Distribution Discrete Probability, 2008.

IJCRT