# **SYSTEM BLOCK DIAGRAM**

1. There are two entities involved: Data Owner (DO) and Security Operation Center (SOC). Security data is owned by the Data owner but lacks the expertise in the field of intrusion detection and thus shares its data with the external SOC which has the required expertise and offers its intrusion detection service to the data owner. DO but it is hesitant to share the data with an external party because of security concerns and do only after having taken all the necessary precautions.

1. First, SOC with the help of an intrusion policy which is just another defined bunch of intrusion detection configurations forms its proprietary detection model.

2. The feature selection process is used to eliminate features which are either redundant or irrelevant to lower the computing time.

3. The data owner then encrypts the security data with its public key using partial homomorphic encryption and sends it to the SOC.

4. After the pattern matching phase, the result of the phase which is encrypted by default is sent to the DO. The DO then decrypts the results using its private key and sends it to the SOC for examination.

5. The SOC then decrypts the result and learns about the offensive records and to which rule in the intrusion policy are these records matched.

6. It then alerts the Data Owner in case of an intrusion and sends the offensive records and also advises on the steps to be taken in case of an attack.

1. The duplicates have already been removed as NSL KDD dataset is already standardized. The nan and Infinity values are replaced with zero initially.
2. DATACLEANING AND PREPROCESSING: Preprocessing operation is done on the dataset as the dataset contains numerical and non-numerical values. One-Hot Encoding is used for this operation. An integer matrix denoting the values of the categorical features is an input to the One Hot Encoder. This will transform all the categorical features to their corresponding binary features out of which one will be active at a time. The dataset is then divided into four parts based on the attacks (U2R, Probe, DoS, R2L) which need to be classified.
3. FEATURE SCALING: Featuring scaling is performed to clear of features which have large values as this will affect the final result. Standard Scaler is used to perform this operation. In Standard Scaler the average for a feature is calculated and then the mean is subtracted from the current value of the feature and the result is divided by the Standard Deviation. The standard deviation will be 1 after each feature is scaled. It is the process in which irrelevant and unnecessary features are eliminated with minimal information loss. Subsets of the features are selected which fully represents all the features in the dataset in terms of accuracy and other metrics. It is also possible that there is a correlation between features when a large number of features are present.
4. Feature selection also helps to eliminate this problem. We have used Recursive Feature Elimination (RFE) to perform this operation. We plot the graph for the accuracy against the number of features and based upon that we select the optimal number of features for each of the attacks. Here, we have built two models: Decision Trees, Random Forest. Both machine learning models are built for all 4 types of attack i.e. U2R, R2L, DoS and Probe. This model is used on the dataset containing every feature (123) and also separately for the features(13) selected after feature selection operation.

7. It is the process in which irrelevant and unnecessary features

are eliminated with minimal information loss. Subsets of the

features are selected which fully represents all the features in

the dataset in terms of accuracy and other metrics. It is also

possible that there is a correlation between features when a

large number of features are present. Feature selection also

helps to eliminate this problem. We have used Recursive

Feature Elimination(RFE) to perform this operation. We plot

the graph for the accuracy against the number of features and

based upon that we select the optimal number of features for

each of the attacks. Here, we have built two models: Decision

Trees, Random Forest. Both machine learning models are built

for all 4 types of attack i.e U2R, R2L, DoS and Probe. This

model is used on the dataset containing every feature(123) and

also separately for the features(13) selected after feature

selection operation.

SCOPE

Also the IDS

generates alerts after a certain time lag as the SOC does not

have any clear information on the output of the intrusion

detection model which is also encrypted and needs to be sent

to the data owner where it is decrypted using the private key.

The decrypted results are then sent to the SOC for analysis and

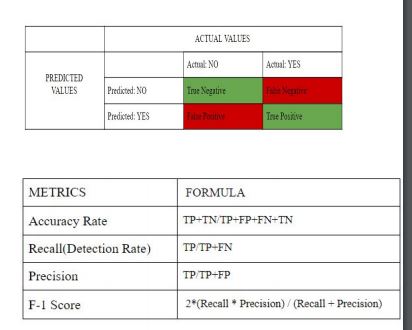
thus the time lag. The Future work would be to try to use

parallel execution to reduce the overhead that comes with

homomorphic encryption and to include other intrusion

detection models and classification methods in our proposed

system.



Denial-of-Service-Attack (DoS): Intrusion where a person aims to make a host inaccessible to its actual purpose by briefly or sometimes permanently disrupting services by flooding the target machine with enormous amounts of requests and hence overloading the host.

User-to-Root-Attack (U2R): Attacker has local access to the victim machine and tries to gain super user privileges.

(R2L): The intrusion in which the attacker can send data packets to the target but has no user account on that machine itself, tries to exploit one vulnerability to obtain local access cloaking themselves as the existing user of the target machine.

Probing-Attack: The type in which the attacker tries to gather information about the computers of the network and the ultimate aim for doing so is to get past the firewall and gaining root access.

CONCLUSION

* We can see that the detection accuracy of an attack is related to the amount of attack data when the model is trained through the experimental results.
* We acknowledge that the solution and analysis of the problem may contain some bias and KDD dataset data could not fully reflect the actual application of the network.
* With the development of network communication, attack behavior will become more and more complex.
* Therefore, collecting, processing and analyzing all aspects of users’ behavior data and applying it to the intrusion detection is important for future development.