JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PROACTIVE APPROACH BASED ANOMALY DETECTION IN NETWORK FORENSICS USING MACHINE LEARNING TECHNIQUE

[1]Vanitha J, [2]Dr. Raja Praveen

[1]M.Tech Student, [2]Associate Professor
[1]M.Tech Cyber Security, FET
[1]Jain (Deemed-to-be University), Bangalore, India
[2] Department of Computer Science and Engineering Faculty of Engineering & Technology
[2]Jain (Deemed-to-be University), Bangalore, India

Abstract: Nowadays, cybercrimes are increasing and have affected large organizations with highly sensitive information. Consequently, the affected organizations spent more resources analyzing the cybercrimes rather than detecting and preventing these crimes. Network forensics plays an important role in investigating cybercrimes; it helps organizations resolve cybercrimes as soon as possible without incurring a significant loss. This paper proposes a novel Proactive Approach to detect anomalies in network using Machine learning Technique. The proposed approach aims to use cybercrime evidence to reconstruct useful attack evidence. Moreover, it helps investigators to resolve cybercrime efficiently. In this paper, the dataset has been taken from real-time network intrusions in a military network environment. The proposed system provides a way to detect and classify multi class anomalies using Random Forest Classifier Algorithm. Random Forest has been selected since it can prevent intrusions up to a good extent by itself and can automatically improve accuracy on anomaly detection. Compared to the reactive approach carried out in network forensics, the results of the proposed machine learning approach prove to be more efficient in terms of time, cost and storage size.

***IndexTerms* – proactive approach, anomaly detection, intrusion detection, machine learning, network forensics.**

I. INTRODUCTION

Current investigation techniques require extensive effort to analyze the overwhelming amount of evidence presented in each cybercrime case which are very costly and time consuming. In addition, gathering useful evidence is difficult because most techniques utilize active and reactive processes to analyze cybercrimes. Such processes start right after the detection of the cybercrime. Network forensic systems can be classified into two approaches: proactive and reactive. Proactive network forensics is a new approach in live investigation that deals with the phases of network forensics during an attack. In contrast, reactive network forensics is a traditional approach that deals with cybercrime cases after a period of time, which consumes a considerable amount of time during the investigation phase. Proactive forensic approaches reduce the time and cost of investigation by identifying potential evidence and reducing the resources needed in the investigation phase. These approaches are utilized in the preliminary analysis of a cybercrime and help improve and accelerate the decision making process.

II. LITERATURE REVIEW

Present network forensics approaches are costly and time consuming. In addition, these approaches normally use active and reactive processes to resolve cybercrimes, and such processes start after the cybercrime has been identified, which makes identifying useful evidence difficult. Moreover, the information required to understand and resolve cybercrime are limited. Mohammad Rasmi et al., proposes a new approach to resolve cybercrime in network forensics. The proposed approach aims to use cybercrime evidence to help investigators to resolve cybercrime efficiently. The paper presents the current network forensics approaches and various existing digital forensics models in order to determine the suitable process to be used in the proposed approach. Thus, the proposed approach based on the generic and modern process model for network forensics. Main phases of network forensics are collection, fusion, identification, examination, correlation, analysis, and documentation of digital evidence. These phases guide other researchers in proposing new approaches for network forensics. The identification of the deliberate intent behind cybercrimes is the main goal of network forensics.

Despite the large volume of research conducted in the field of intrusion detection, finding a perfect solution of intrusion detection systems for critical applications is still a major challenge. This is mainly due to the continuous emergence of security threats which can bypass the outdated intrusion detection systems. The main objective of the paper proposed by Setareh Roshan et.al., is an adaptive design of intrusion detection systems on the basis of Extreme Learning Machines. The proposed system offers the capability of detecting known and novel attacks and being updated according to new trends of data patterns provided by security experts in a cost-effective manner. In this paper, they addressed the problem of adaptability in the field of intrusion detection by proposing a new intrusion detection system. The proposed IDS is an adaptive solution which provides the capability of detecting known and novel attacks as well as being updated according to the new input from human experts in a cost-effective manner. Two novel approaches were presented for updating the system according to the new available information with a low computational cost. These approaches can be used for cases where a human expert requests for modifying the cluster assignment of existing data or a new class of data is available as the input. In these cases, the proposed approaches update the model without performing a full retraining.

Abhishek Verma et.al., basically used two techniques, k-means clustering and k-nearest neighbor classification to measure the complexity in terms of prominent metrics. It deals with the evaluation and the statistical analysis of labeled flow based CIDDS-001 dataset used for evaluating Anomaly based Network Intrusion Detection Systems. Based on evaluation, they concluded that both k-means clustering k-nearest neighbor classification perform well over CIDDS-001 dataset in terms of used prominent metrics. Hence the dataset can be used for the evaluation of Anomaly based Network Intrusion Detection Systems.

The Internet and computer networks are exposed to an increasing number of security threats. With new types of attacks appearing continually, developing flexible and adaptive security oriented approaches is a severe challenge. In this context, anomaly-based network intrusion detection techniques are a valuable technology to protect target systems and networks against malicious activities. However, despite the variety of such methods described in the literature in recent years, security tools incorporating anomaly detection functionalities are just starting to appear, and several important problems remain to be solved. Garcıa-Teodoro et.al., begins with a review of the most well-known anomaly-based intrusion detection techniques. Then, available platforms, systems under development and research projects in the area are presented. Finally, they outline the main challenges to be dealt with for the wide scale deployment of anomaly-based intrusion detectors, with special emphasis on assessment issues.

Mahdi Zamani reviewed several influential algorithms for intrusion detection based on various machine learning techniques. Characteristics of ML techniques makes it possible to design IDS that have high detection rates and low false positive rates while the system quickly adapts itself to changing malicious behaviors. Division of these algorithms into two types of ML-based schemes: Artificial Intelligence (AI) and Computational Intelligence (CI). Although these two categories of algorithms share many similarities, several features of CI-based techniques, such as adaptation, fault tolerance, high computational speed and error resilience in the face of noisy information, conform the requirement of building efficient intrusion detection system.

In recent years, wireless ad hoc sensor network becomes popular both in civil and military jobs. However, security is one of the significant challenges for sensor network because of their deployment in open and unprotected environment. As cryptographic mechanism is not enough to protect sensor network from external attacks, intrusion detection system needs to be introduced. Though intrusion prevention mechanism is one of the major and efficient methods against attacks, but there might be some attacks for which prevention method is not known. Besides preventing the system from some known attacks, intrusion detection system gather necessary information related to attack technique and help in the development of intrusion prevention system. In addition to reviewing the present attacks available in wireless sensor network this paper examines the current efforts to intrusion detection system against wireless sensor network. Mohammad Saiful Islam Mamun et.al., proposes a hierarchical architectural design based intrusion detection system that fits the current demands and restrictions of wireless ad hoc sensor network. In this proposed intrusion detection system architecture we followed clustering mechanism to build a four level hierarchical network which enhances network scalability to large geographical area and use both anomaly and misuse detection techniques for intrusion detection. They introduce policy based detection mechanism as well as intrusion response together with GSM cell concept for intrusion detection architecture.

III. Proposed System

Sensitive and dynamic information constantly flowing between devices calls for proactive action. A proactive approach to cyber security includes preemptively identifying security weaknesses and adding processes to identify threats before they occur. For the proactive event detection approach, we use the cyber security dataset that is publically available and it is mostly used for evaluating Intrusion Detection System in networks. The system proposed is composed of Preprocessing, Feature selection and Machine Learning algorithm as shown in Fig.1. Data preprocessing plays an important role to make the data ready for the prediction process. In data preprocessing, the transformation and normalization operation is performed on the dataset. It can help to better expose the underlying structure of the data to the learning algorithm and, in turn, may result in better predictive performance. Feature selection component are responsible to extract most relevant features or attributes to identify the instance to a particular group or class. The Machine learning algorithm such as Random Forest is applied in order to detect hacks and data breaches. A system is developed to detect all possible attacks signatures coming from malicious user's request and then generate a training model. One has to train the network to recognize various normal and abnormal traffic behavior. After training, we use a test dataset to evaluate the data. In realtime, upon receiving new request the system will apply that request on that train model to predict it class whether request belongs to normal class or anamoly class. The classifying of datasets based on the anomaly class aids in predicting the vulnerability of individual attacks in various networks.

IV. METHODOLOGY

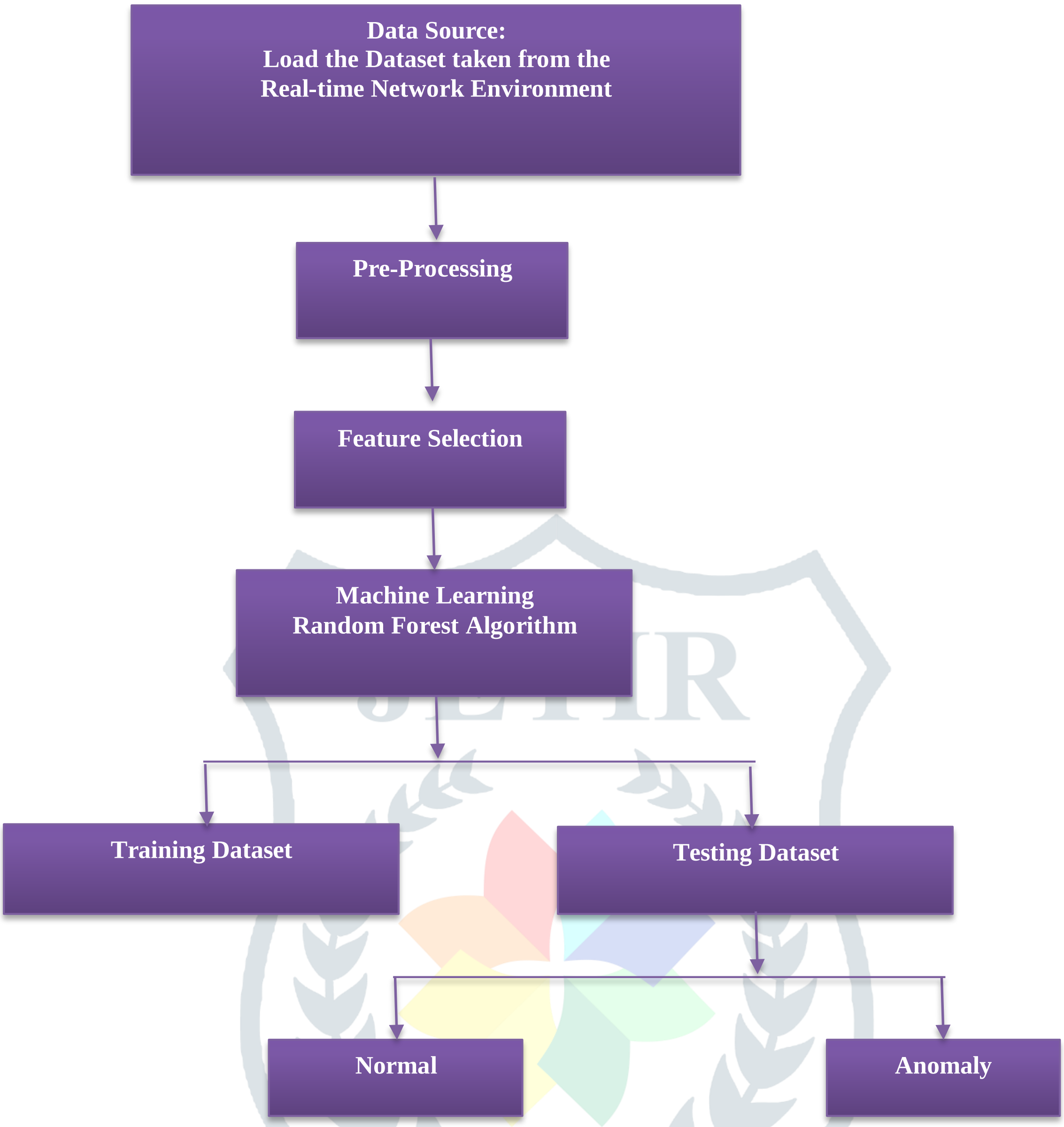


Fig.1. Proposed Methodology

V. MODULES OF THE PROPOSED SYSTEM

The proposed system contains the following modules:

1. Dataset
2. Data Preprocessing
3. Feature Selection
4. Build the Model
5. Model Prediction and Test Results

5.1 Dataset

The dataset has been taken from the publicly available dataset online from the following link:https://kdd.org/kdd-cup/view/kdd-cup-1999/Data. This dataset contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Software to detect network intrusions protects a computer network from unauthorized users, including perhaps insiders. The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between bad connections, called intrusions or attacks, and good as normal connections.

The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, was provided. The 1999 KDD intrusion detection contest uses a version of this dataset.

Lincoln Labs set up an environment to acquire nine weeks of raw TCP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. They operated the LAN as if it were a true Air Force environment, but peppered it with multiple attacks.

The raw training data was about four gigabytes of compressed binary TCP dump data from seven weeks of network traffic. This was processed into about five million connection records. Similarly, the two weeks of test data yielded around two million connection records.

A connection is a sequence of TCP packets starting and ending at some well-defined times, between which data flows to and from a source IP address to a target IP address under some well-defined protocol. Each connection is labeled as either normal, or as an attack, with exactly one specific attack type. Each connection record consists of about 100 bytes.

Attacks fall into four main categories:

DOS: denial-of-service, e.g. syn flood;
R2L: unauthorized access from a remote machine, e.g. guessing password;
U2R: unauthorized access to local superuser (root) privileges, e.g., various ``buffer overflow'' attacks;
Probing: surveillance and other probing, e.g., port scanning.

It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data. This makes the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the "signature" of known attacks can be sufficient to catch novel variants. The datasets contain a total of 24 training attack types, with an additional 14 types in the test data only.

5.2 Data Preprocessing

One hot encoding is a technique used to represent categorical variables as numerical values in a machine learning model. The advantages of using one hot encoding include:

- ➢ It allows the use of categorical variables in models that require numerical input.
- ➢ It can improve model performance by providing more information to the model about the categorical variable.

In our project,the categorical values are replaced by numerical values.

5.3 Feature Selection

Feature extraction involves selecting specific relevant features to create the training and testing datasets that would be fed into the algorithm. This achieves a couple of aims; it reduces the probability of overfitting, it makes interpretation easier and it improves the chance for generalization.

In this module we eliminate redundant and irrelevant data by selecting a subset of relevant features that fully represents the given problem. This analyzes each feature individually to determine the strength of the relationship between the feature and labels.

5.3 Build the Model

After Preprocessing and Feature Selection is completed, here we split the dataset into training and testing. We train the model using Random Forest Algorithm.

5.4 Model Prediction and Test Results

The evaluation is carried out with the test data to make predictions of the model such as normal or malicious attack. These tests were conducted using a normal train/test split

VI. Results And Discussions

6.1 GUI Design

The GUI was designed for the user interface. Fig.2 shows the result of the GUI design.

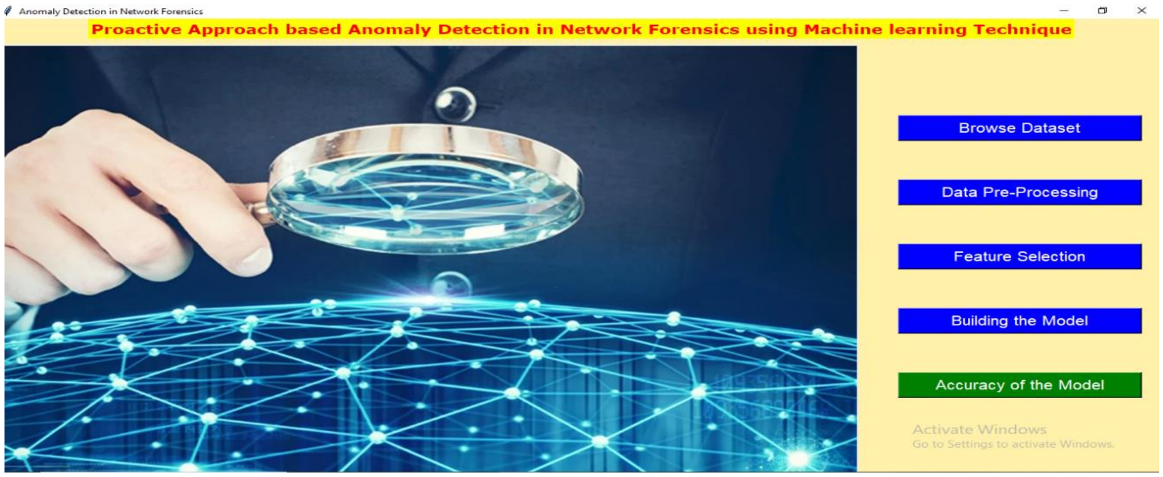

Anomaly Detection in Network Forensics
Proactive Approach based Anomaly Detection in Network Forensics using Machine learning Technique
Browse Dataset
Data Pre-Processing
Feature Selection
Building the Model
Accuracy of the Model
Activate Windows
Go to Settings to activate Windows.

Fig.2. GUI Design

6.2 Loading the Dataset

The dataset is loaded using the browse button provided in the GUI Design as shown in the Fig.3 and Fig.4 shows the evidence file that has been loaded.

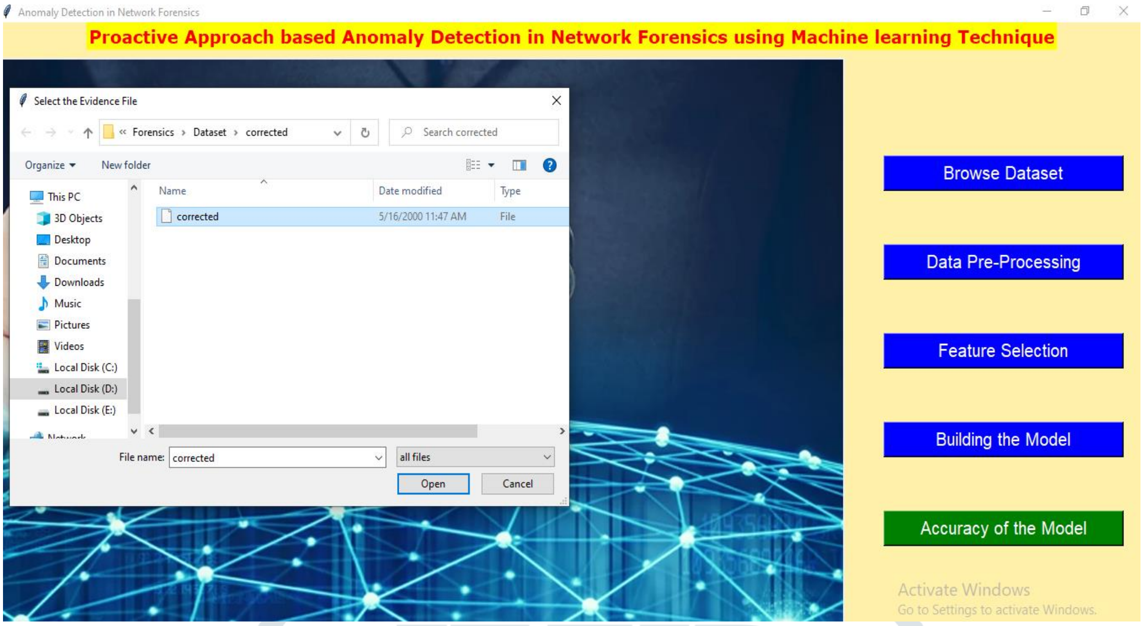
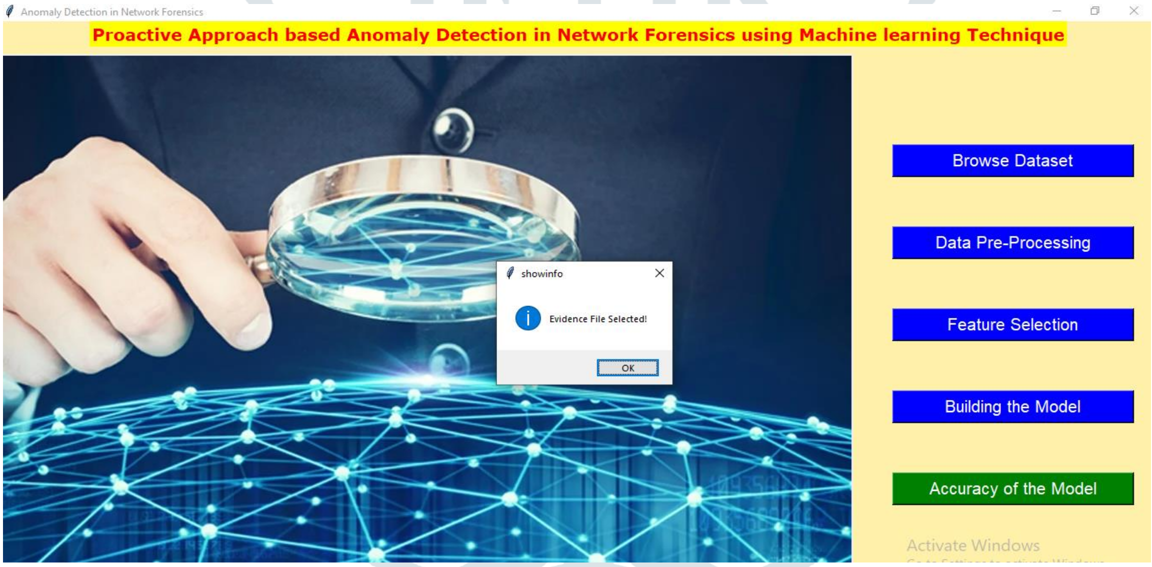


Fig.3. Loading the Dataset

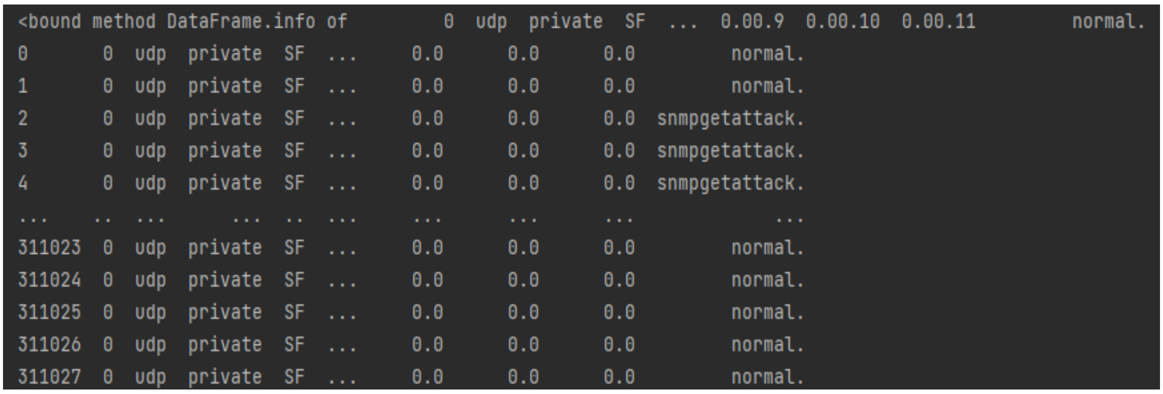

Anomaly Detection in Network Forensics
Proactive Approach based Anomaly Detection in Network Forensics using Machine learning Technique
showinfo
OK
Browse Dataset
Data Pre-Processing
Feature Selection
Building the Model
Accuracy of the Model
Activate Windows

Evidence File Selected!

Fig.4. Evidence File Selection

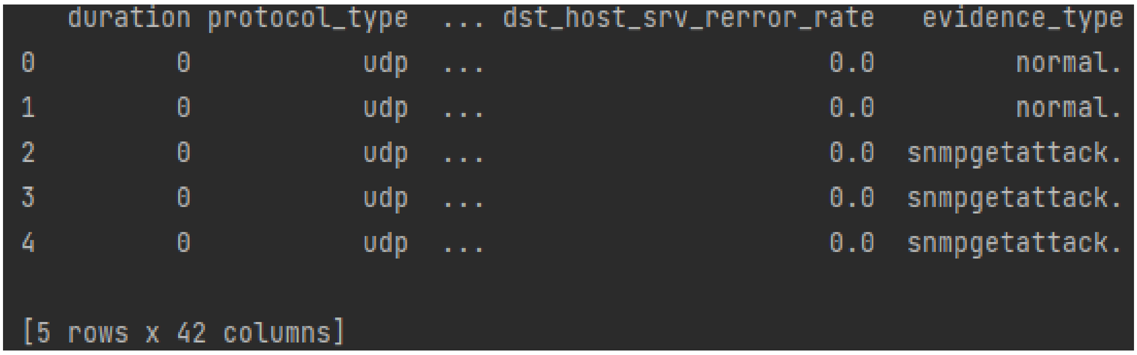
6.3 Preprocessing the Dataset

After the raw dataset is loaded as shown in the Fig 5, the unlabeled dataset is converted to labeled dataset as shown in the Fig.6. One-Hot-Encoding method is used to convert categorical features to binary values. Then we distribute categories in services. Now, the features are transformed using Label Encoder, to transform every category to a number as shown in the Fig.7. After that we split the dataset into 4 sub categories each of separate attack type. Here, we provide a dummy name to each attack type as 0=normal, 1=DOS, 2=R2L, 3=U2R, 4=probe. After this feature scaling is done by splitting features into X and Y where X as a data frame of features and Y as a series of output variables.



```
<bound method DataFrame.info of        0  udp  private  SF  ...  0.00.9  0.00.10  0.00.11        normal.
0       0  udp  private  SF  ...     0.0      0.0      0.0        normal.
1       0  udp  private  SF  ...     0.0      0.0      0.0        normal.
2       0  udp  private  SF  ...     0.0      0.0      0.0  snmpgetattack.
3       0  udp  private  SF  ...     0.0      0.0      0.0  snmpgetattack.
4       0  udp  private  SF  ...     0.0      0.0      0.0  snmpgetattack.
...    ..  ...      ...  ..  ...     ...      ...      ...            ...
311023  0  udp  private  SF  ...     0.0      0.0      0.0        normal.
311024  0  udp  private  SF  ...     0.0      0.0      0.0        normal.
311025  0  udp  private  SF  ...     0.0      0.0      0.0        normal.
311026  0  udp  private  SF  ...     0.0      0.0      0.0        normal.
311027  0  udp  private  SF  ...     0.0      0.0      0.0        normal.
```

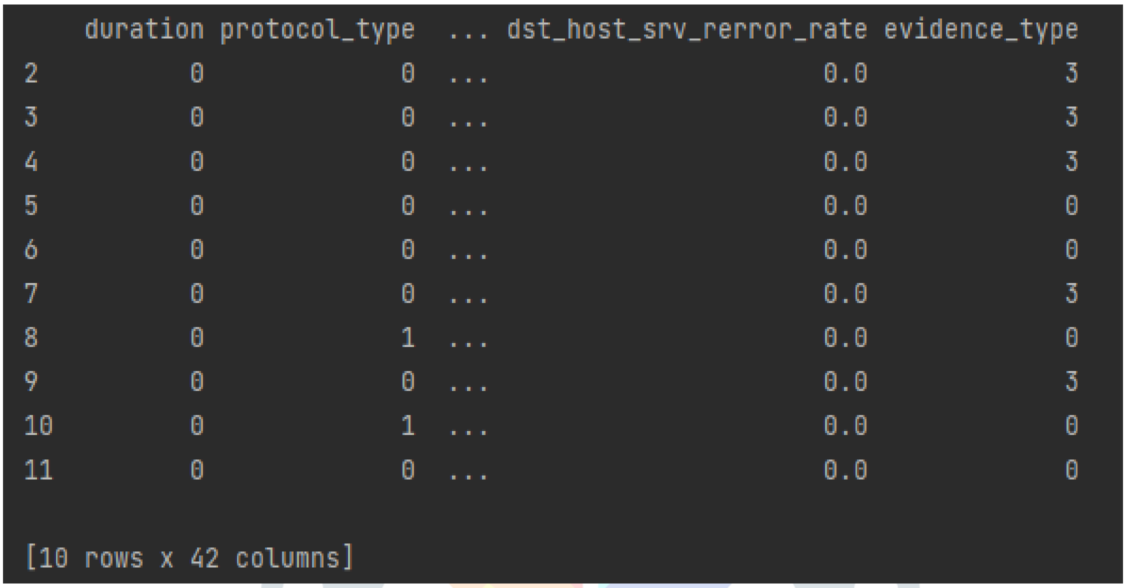
Fig.5. Raw Dataset



```
   duration protocol_type  ... dst_host_srv_rerror_rate   evidence_type
0         0           udp  ...                      0.0         normal.
1         0           udp  ...                      0.0         normal.
2         0           udp  ...                      0.0  snmpgetattack.
3         0           udp  ...                      0.0  snmpgetattack.
4         0           udp  ...                      0.0  snmpgetattack.

[5 rows x 42 columns]
```

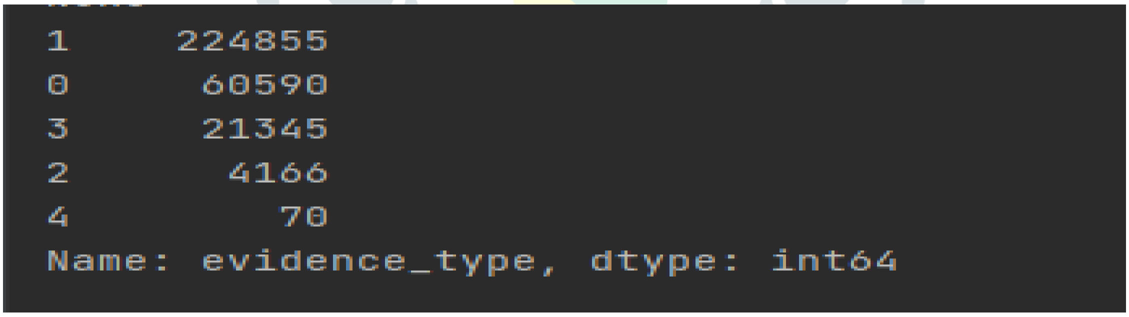
Fig.6. Labeled Dataset



```
    duration protocol_type  ... dst_host_srv_rerror_rate evidence_type
2          0             0  ...                      0.0             3
3          0             0  ...                      0.0             3
4          0             0  ...                      0.0             3
5          0             0  ...                      0.0             0
6          0             0  ...                      0.0             0
7          0             0  ...                      0.0             3
8          0             1  ...                      0.0             0
9          0             0  ...                      0.0             3
10         0             1  ...                      0.0             0
11         0             0  ...                      0.0             0

[10 rows x 42 columns]
```

Fig.7. Label Encoder



```
1     224855
0      60590
3      21345
2       4166
4         70
Name: evidence_type, dtype: int64
```

Fig.8. Labeled Dataset

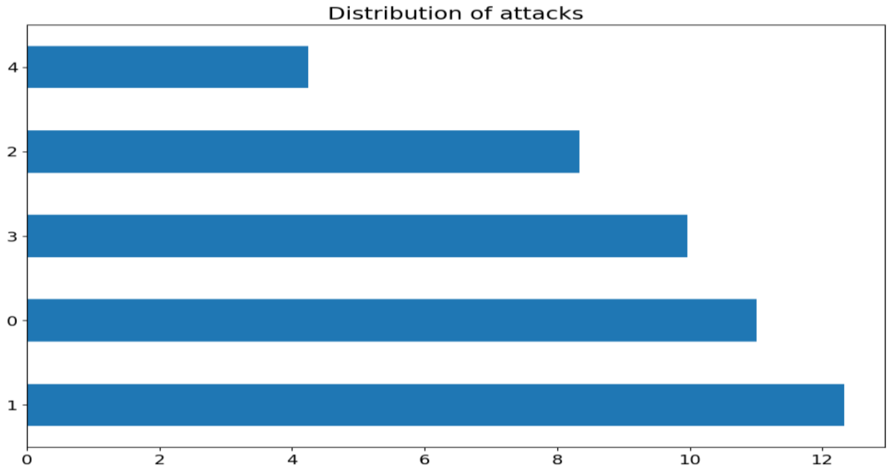


Fig.9. Distribution of Attacks

6.4 Selecting the Features

The features are selected using the Feature selection button. Feature selection step is correlated to data pre-processing step where irrelevant features are removed which increases the accuracy. Feature Selection refers to identification of features that are strongly correlated to the problem which are useful in the prediction of a class. Recursive Feature Elimination, or RFE for short, is a popular feature selection algorithm that has been chosen. RFE is popular because it is easy to configure and use and because it is effective at selecting those features (columns) in a training dataset that are more or most relevant in predicting the target variable. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains as shown in the Fig.10..In this project, Random Forest Algorithm is used for fitting purpose. After selecting relevant features ranking is done to provide priority to features for better accuracy. This process ends with the message training data feature selection completed as shown in the Fig.11.

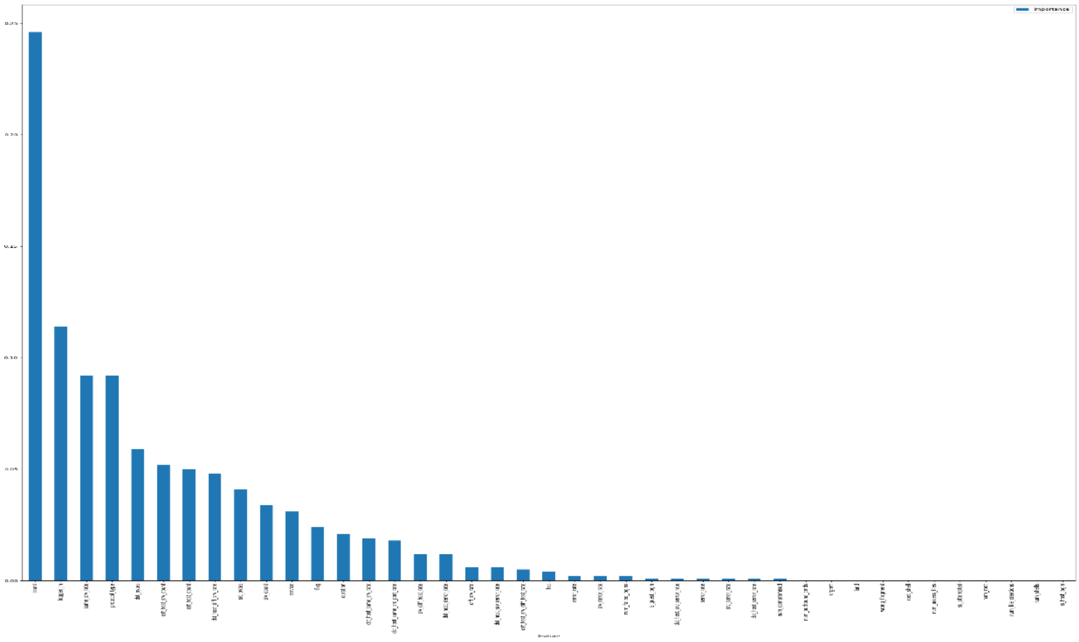
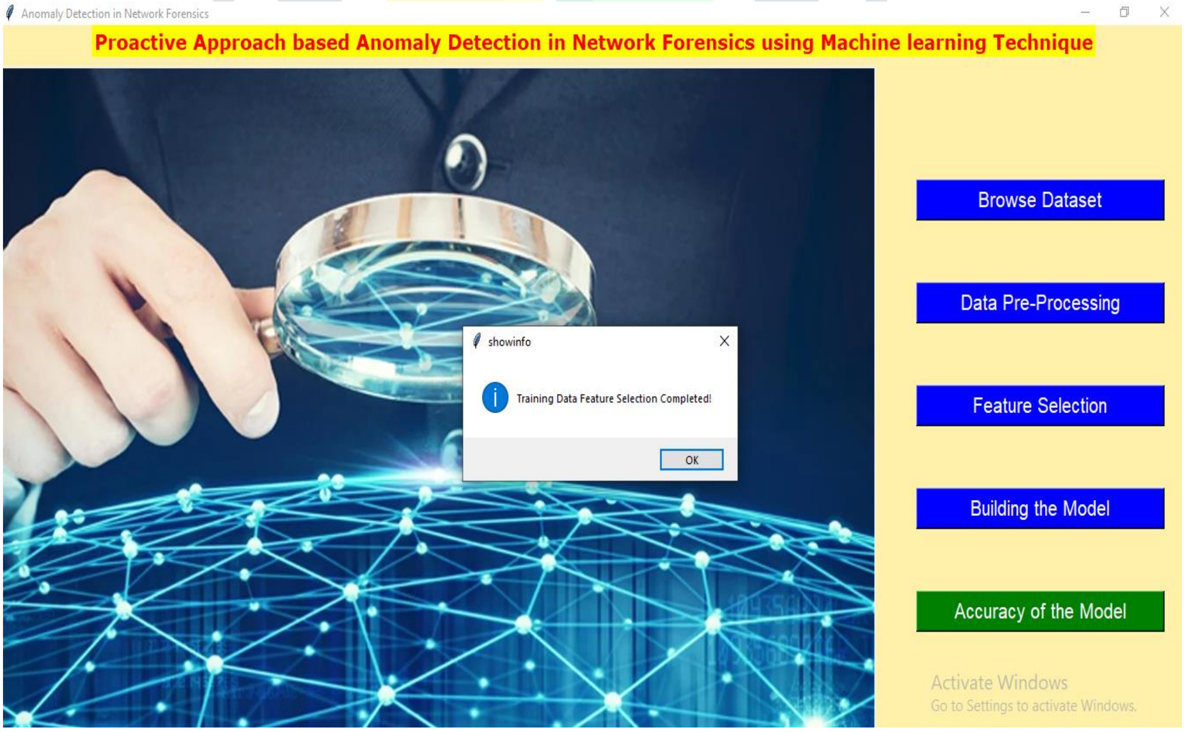


Fig.10.Feature Selection



Proactive Approach based Anomaly Detection in Network Forensics using Machine learning Technique

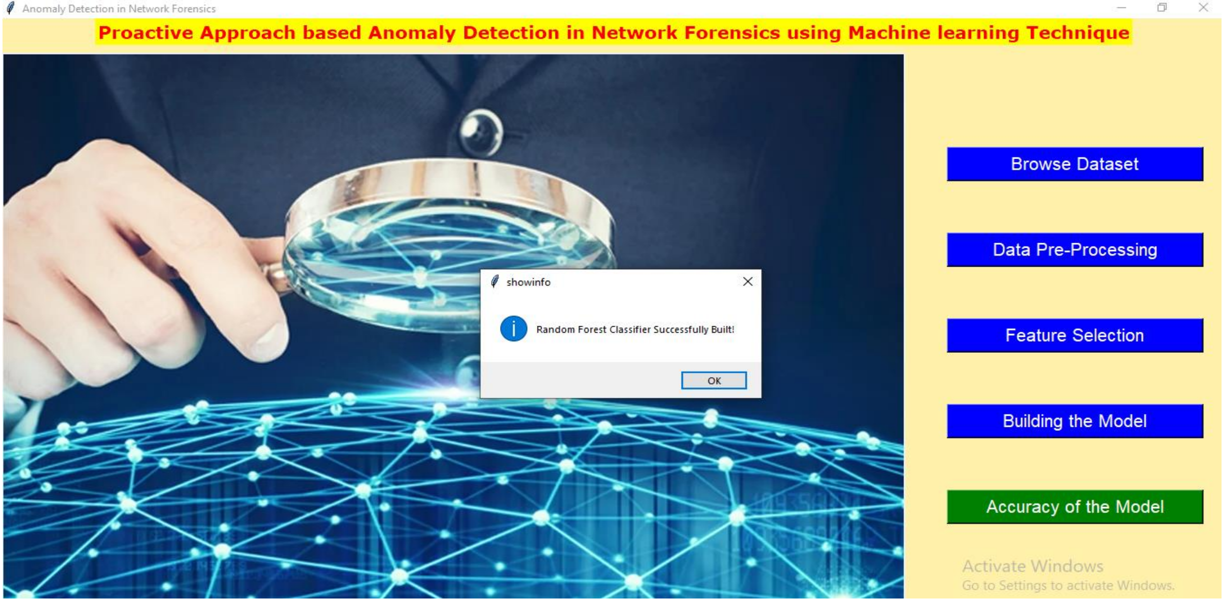
showinfo

Training Data Feature Selection Completed!

Fig.11. GUI design- Feature Selection

6.5 Building the Model

To get a high accurate classifier which deals with real time data, a high performing model selection is required. In this paper, an efficient model is built using the Random Forest(RF) classifier algorithm as shown in the Fig.12. Here, classifiers are trained for all features using trained dataset. To perform this model classification the predefined function of python called **RandomForestClassifier()** has been used. Here, classification model is built for all types of attack.



Anomaly Detection in Network Forensics

Proactive Approach based Anomaly Detection in Network Forensics using Machine learning Technique

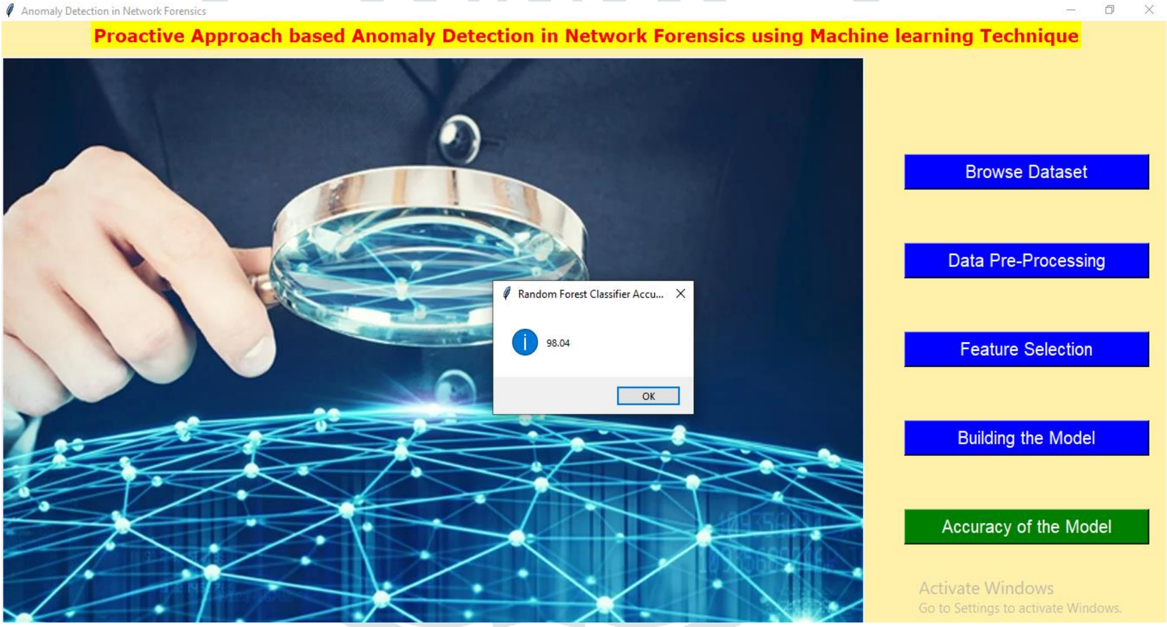
showinfo

Random Forest Classifier Successfully Built!

Fig.12. Building the Model

6.6 Test Results of the Model

The accuracy achieved by the random forest classifier is shown in Fig.13. below.



Anomaly Detection in Network Forensics

Proactive Approach based Anomaly Detection in Network Forensics using Machine learning Technique

Random Forest Classifier Accu...

98.04

Fig.13. Accuracy of the Model

The dataset for analysis consists of 42 attributes where last attribute has class labels. To evaluate the performance of classifier we construct confusion matrix to find accuracy, precision, recall and f-measure.The obtained confusion matrix is shown below in the Fig.14.

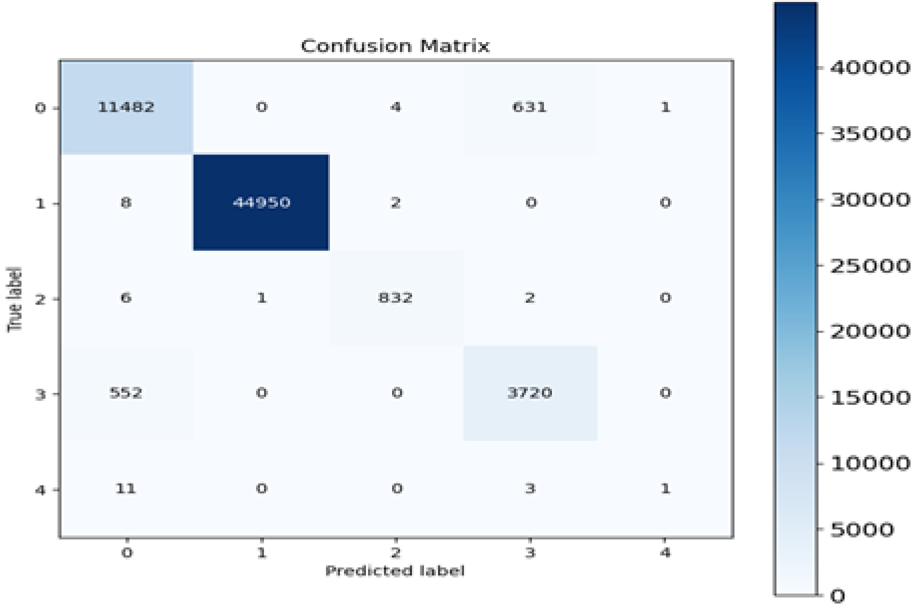


Fig.14. Confusion Matrix

VII. Conclusion

This project deals with Proactive Approach Based Anomaly Detection using Random Forest(RF) algorithm to detect four types of attack that are DOS, Probe, R2L, U2R by reducing the features. The main motive of our feature selection process on dataset is to reduce dimensionality, to remove redundancy of data and removal of irrelevant features. Recursive Feature Elimination along with Random Forest (RF) algorithm is applied for Feature Selection. The obtained result shows that Random Forest classification with reduced features is more accurate than that of found from all features. Our experimental result shows that accuracy, precision, recall, f-measure of these four attack types has increased by our proposed methodology

References

[1] Mohammad Rasmi, Aman Jantan, Hani Al-Mimi, "A New Approach For Resolving Cyber Crime In Network Forensics Based On Generic Process Model",ICIT 2013 The 6th International Conference on Information Technology.

[2] Setareh Roshana,Yoan Michec, Anton Akusokd, Amaury Lendasse, "Adaptive and online network intrusion detection system using clustering and Extreme Learning Machines" - in Journal of the Franklin Institute Volume 355, Issue 4, March 2018, Pages 1752-1779

[3] Abhishek Verma and Virender Ranga,"Statistical analysis of CIDDS-001 dataset for Network Intrusion Detection Systems using Distance-based Machine Learning" - in 6th International Conference on Smart Computing and Communications,10.1016/j.procs.2017.12.09

[4] P. Garcı´a-Teodoroa, J. Dı´az-Verdejoa, G. Macia´-Ferna´ndeza, E. Va´zquez,"Anomaly-based network intrusion detection: Techniques, systems and challenges" - in computers & security 28 (2009) 18–28,doi:10.1016/j.cose.2008.08.003.

[5] Mohammad Saiful Islam Mamun and A.F.M. Sultanul Kabir,"Hierarchical Design Based Intrusion Detection System For Wireless Ad Hoc Sensor Network" - in International Journal of Network Security & Its Applications (IJNSA), Vol.2, No.3, July 2010