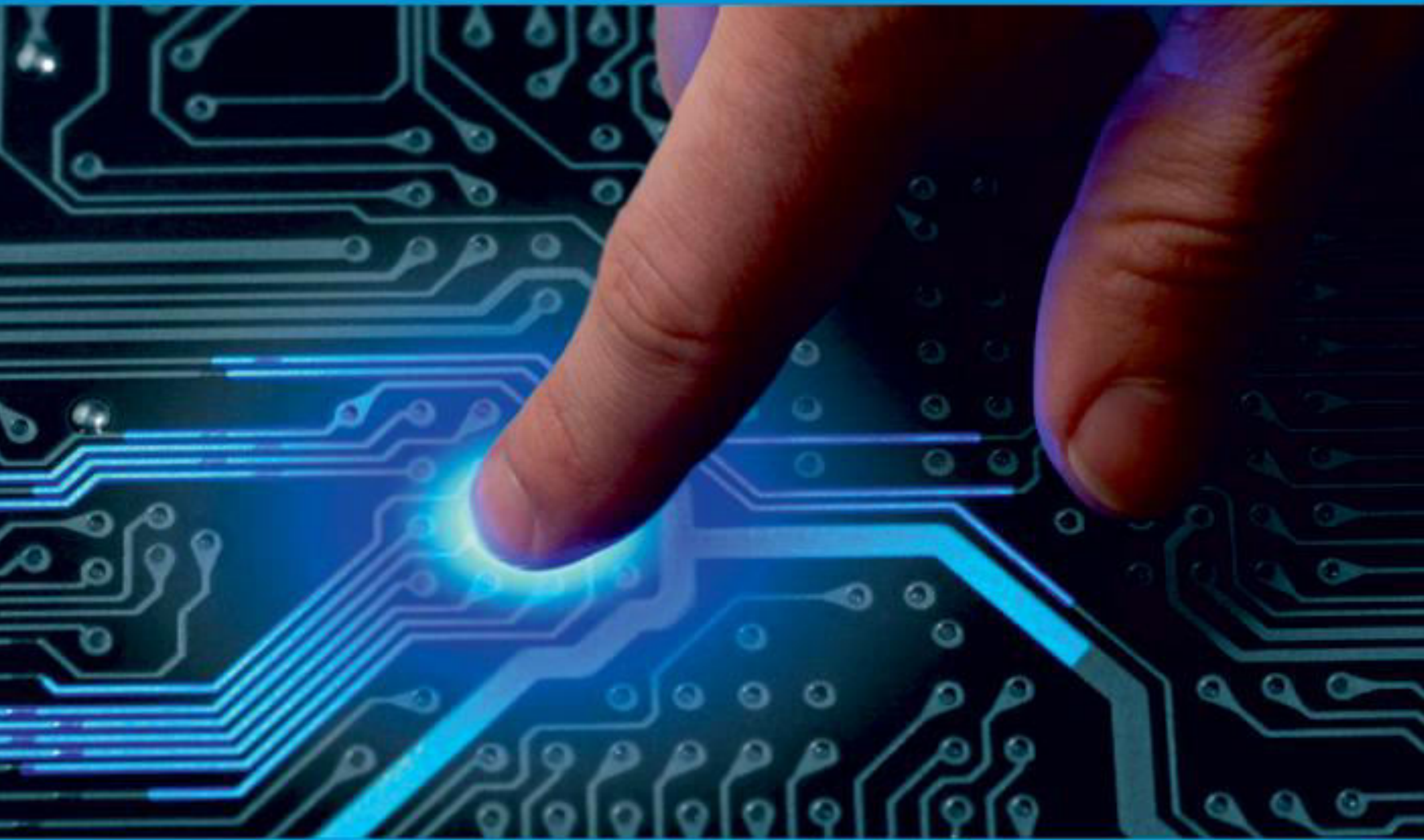
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Machine Learning Approach for Anomaly Detection of IoT Cyber Attacks

M. Selvarani, K.Kuralamudhu, M. Sakthi Sinega, J. Sangavi, R. Santhiya

Final year ECE, Sir Issac Newton College of Engg. & Technology, Nagapattinam, Tamil Nadu, India

Assistant Professor, Sir Issac Newton College of Engg. & Technology, Nagapattinam, Tamil Nadu, India

ABSTRACT: With the evolution in wireless communication, there are many security threats over the internet. The intrusion detection system (IDS) helps to find the attacks on the system and the intruders are detected. Previously various machine learning (ML) techniques areapplied on the IDS and tried to improve the results on the detection of intruders and to increase the accuracy of the IDS. This paperhas proposed an approach to develop efficient IDS by using the principal component analysis (PCA) and the random forestclassification algorithm. Where the PCA will help to organize the dataset by reducing the dimensionality of the dataset and the random forest will help in classification. Results obtained states that the proposed approach works more efficiently in terms of accuracy as compared to other techniques like SVM, Naive Bayes, and Decision Tree. The results obtained by proposed method are having the values for performance time (min) is 3.24 minutes,Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21 %.

KEYWORDS: Machine Learning, Dataset, Anomaly Detection,Security, Cyber Attacks, IoT Security.

I. INTRODUCTION

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificialIntelligence which combines data with statistical tools to predict an output which can be used to make actionable insights.Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation.

Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similarexample, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.

The machine uses some fancy algorithms to simplify the reality and transform this discovery into a model. Therefore, the learning stage isused to describe the data and summarize it into a model.For instance, the machine is trying to understand the relationship between thewage of an individual and the likelihood to go to a fancy restaurant.

It turns out the machine findsa positive relationship between wage and going to a high-end restaurant: This is the model

Inferring

When the model is built, it is possible to test how powerful it is on never-seen- before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. Incases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers. For example, to train a system for the task of digital character recognition, the MNIST dataset of handwritten digits has often been used.

II. Related work

Smart home devices have brought in a disruptive, revolutionary Internet-based ecosystem that enhanced our daily lives but has pushed private data from inside our homes to external public sources. Threats and attacks mounted against IoT deployments have only increased in recent times.There have been several proposals to Secure home automation environments, but there is no full protection against Cybersecurity threats for our home IoT platforms.This research investigates attack attempts on smart home environments, focusing on firmware, brute force, and DoS attacks on the Internet of Things (IoT) network which were successful in bringing down the device in less than a minute. Weak passwords were cracked usingBrute Force techniques related to HTTP, SSH, Telnet, and FTP protocols, and an unknown service port to reveal backdoor access. Cross-site scripting vulner- ability was detected on IoT devices that could allow running malicious scripts on the devices. The authorsalso exploited the unknown services to reveal backdoors and access sensitive device details and potentially exploited them to add new ports or rules to turn the IoT devices into a router to attack other devices. To detect and mitigate such attacks, the authors present an IoT-based intrusion detection and prevention system to secure smart home network devices. The authors compared the proposed framework with other similar research based on Precision, Accuracy, F-measure, and Recall. The proposed model outperforms all the other known models reporting a high of 95% for identifying maliciousAttack packets, while others reported 58% and 71% detection percentage.

III. Materials

- HARDWARE REQUIREMENTS

Selection System: Pentium IV 2.4 GHz.Hard Disk: 40 GB.
Floppy Drive 1.44 Mb.

Monitor: 15 VGA Colour.Mouse: Logitech.
Ram: 512 Mb.

- SOFTWARE REQUIREMENTS:

Operating system: Windows 7.Coding Language: Python
Database: MYSQL.

IV. Methodology

Image PCA (Principal Component Analysis) and Random Forest are two popular machine learning algorithms that can be used for anomaly detection in self-organizing networks. Here is a comparison of using these algorithms for conventional versus contemporary machine learning:

1. Conventional Machine Learning: In conventional machinelearning, PCA and Random Forest can be used as follows:

PCA: PCA can be used to reduce the dimensionality of the input databy projecting it onto a lower-dimensional subspace. The resulting projection can then be used as input to a conventional machine learning algorithm, such as a support vector machine (SVM) or logistic regression, for anomaly detection.

Random Forest: Random Forest is an ensemble learning algorithm that can be used for both classification and regression problems. In the case of anomaly detection, it can be used as a classifier to detect anomalies based on the input data.

2. Contemporary Machine Learning: In contemporary machinelearning, PCA and Random Forest can be used as follows:

PCA: In contemporary machine learning, PCA can be used as apreprocessing step to automatically learn the features from the input data. The resulting features can then be used as input to a contemporary machine learning algorithm, such as a deep neuralnetwork or auto encoder, for anomaly detection.

Random Forest: Random Forest can be used in contemporary machine learning as an ensemble method to combine the outputs of multiple contemporary machine learning models, such as deep neuralnetworks or auto encoders, to detect anomalies.

The key difference between using PCA and Random Forest in conventional versus contemporary machine learning is in their usage.In conventional machine learning, they are used as standalone algorithms or in combination with other conventional machine learning algorithms. In contrast, in contemporary machine learning, they are used in combination with other contemporary machine learning algorithms, such as deep neural networks or auto encoders.

V. HOW TO TRAIN THE MODEL?

Training a model for anomaly detection of IoT cyber-attacks involvesseveral steps:

1. Data Collection: Gather a diverse dataset of IoT network traffic or logs that include both normal and attack scenarios.

2. Preprocessing: Clean the data, handle missing values, and transform features if needed. This step may also involve featureengineering to extract meaningful information from the data.

3. Feature Selection: Identify relevant features that contribute most to distinguishing between normal and anomalous behavior. This could involve techniques like statistical analysis, correlation analysis, or domain knowledge.

4. Model Selection: Choose a suitable anomaly detection algorithm. Common approaches include Isolation Forest, One-Class SVM, k-means clustering, and auto encoders for deep learning-basedapproaches.

5. Training: Train the selected model on the labeled dataset. Ensure to split the dataset into training and validation sets to evaluate the model's performance effectively.

6. Evaluation: Assess the model's performance using appropriateevaluation metrics such as precision, recall, F1-score, ROC curve, and AUC-ROC.

7. Fine-tuning: Adjust model hyper parameters or try different algorithms to improve performance if necessary. This step may involve techniques like cross-validation or grid search.

8. Deployment: Once satisfied with the model's performance, deploy it to monitor real-time IoT network traffic for anomalies. Ensure to continuously monitor and update the model as new attack patterns emerge.

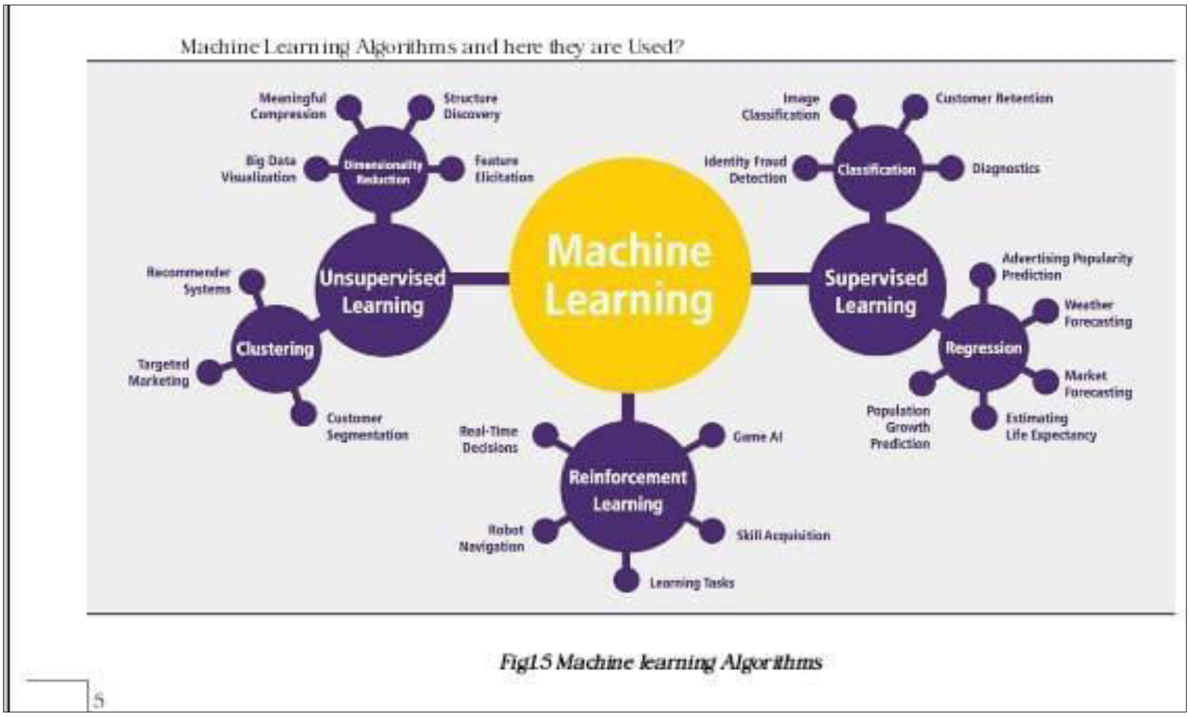
9. Monitoring and Maintenance: Regularly monitor the deployedmodel's performance and update it as needed to adapt to evolving cyber threats and changes in the IoT environment.

10. Documentation: Document the entire process, including data sources, preprocessing steps, model selection, training parameters, and evaluation results, for future reference and reproducibility.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

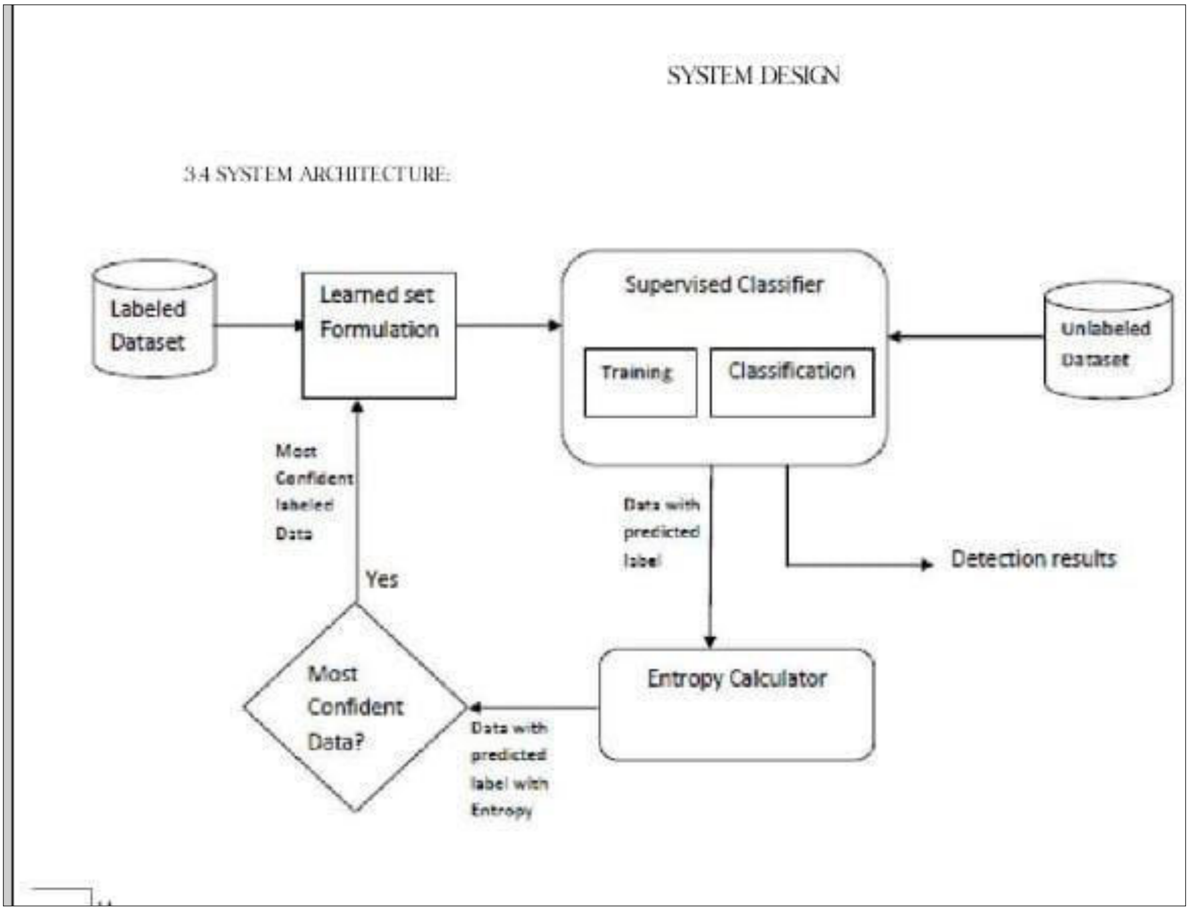
| DOI: 10.15680/IJIRCCE.2023.1205076 |



Machine Learning Algorithms and here they are Used?

Fig1.5 Machine learning Algorithms

5



SYSTEM DESIGN

3.4 SYSTEM ARCHITECTURE:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

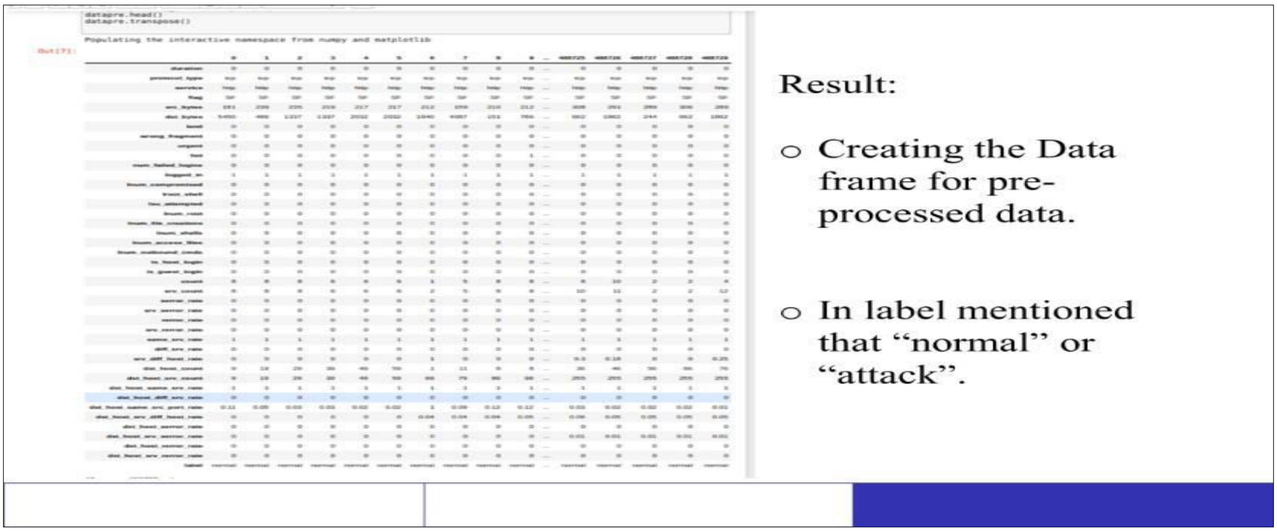
| DOI: 10.15680/IJIRCCE.2023.1205076 |

VI. RESULTS AND DISCUSSIONS

Conventional Machine Learning: In general, conventional machine learning algorithms, such as PCA and Random Forest, can achieve good accuracy in detecting anomalies in self-organizing networks when the data is well-defined and has a clear separation between normal and anomalous data. However, their accuracy may decrease when the data is complex and has subtle differences between normal and anomalous data. In such cases, the algorithms may require additional feature engineering or the use of more complex models.

Contemporary Machine Learning: Contemporary machine learning algorithms, such as deep neural networks and autoencoders, have shown promising results in detecting anomalies in self-organizing networks. These algorithms can automatically learn relevant features from the input data, which can capture complex relationships between different variables. Additionally, these algorithms can handle high- dimensional and complex data, which may not be possible with conventional machine learning algorithms. However, the accuracy of contemporary machine learning algorithms can be influenced by factors such as the choice of architecture, the numberof training examples, and the quality of the training data.

The results obtained by proposed method are having the values forperformance time (min) is 3.24 minutes, Accuracy rate (%) is 96.78 %, and the Error rate (%) is 0.21 %....



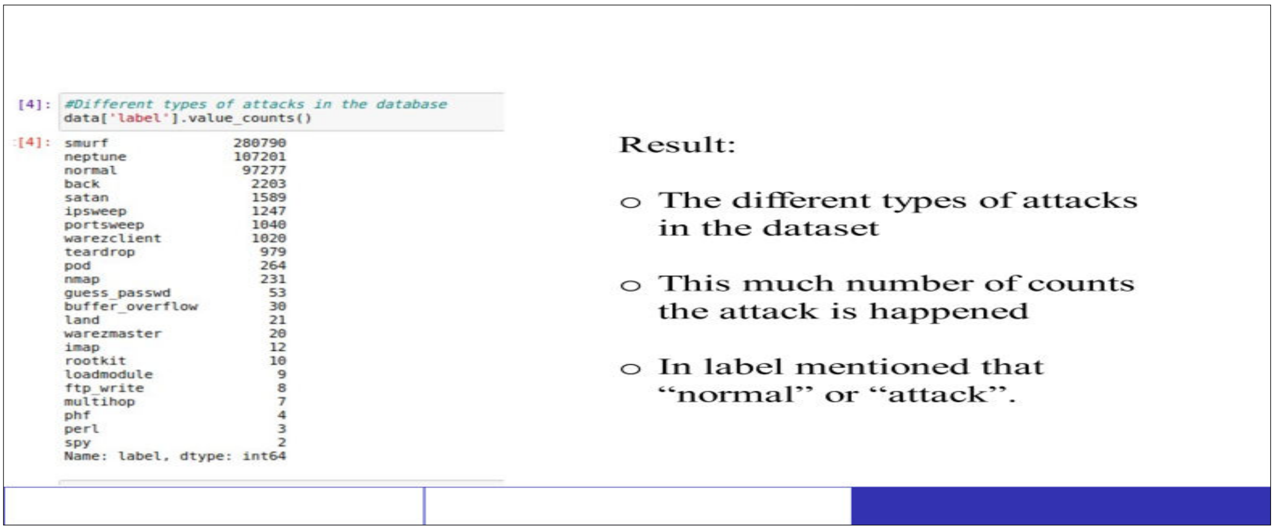
```
datapre.head()
datapre.transpose()
```

```
Populating the interactive namespace from numpy and matplotlib
```

Out[7]:

Result:

- Creating the Data frame for pre-processed data.
- In label mentioned that "normal" or "attack".

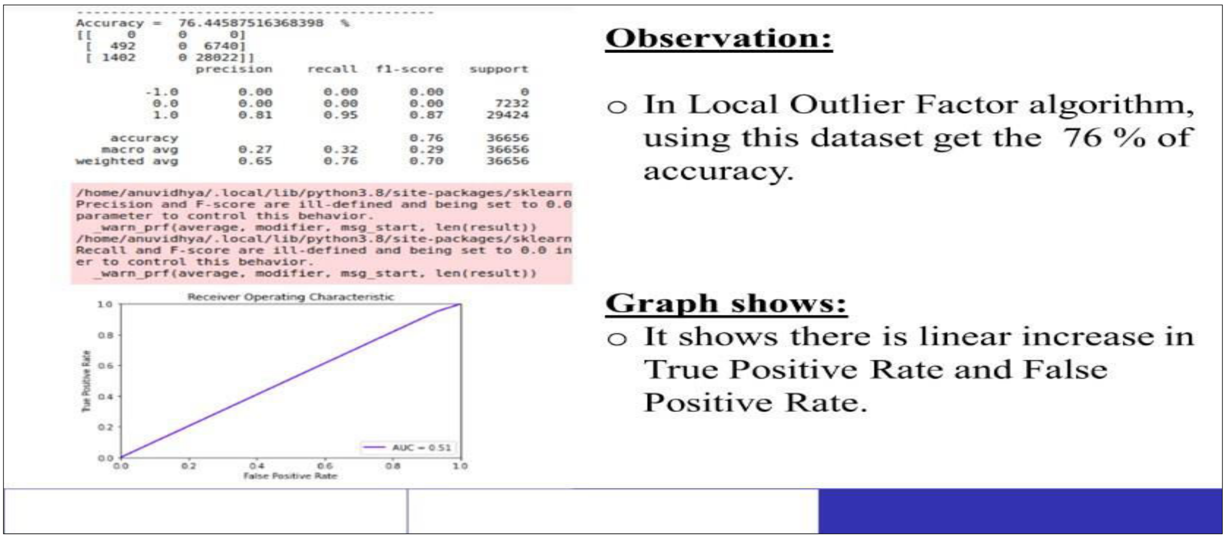


```
[4]: #Different types of attacks in the database
     data['label'].value_counts()
```

```
[4]: smurf             280790
     neptune           107201
     normal             97277
     back                2203
     satan               1589
     ipsweep             1247
     portsweep           1040
     warezclient         1020
     teardrop             979
     pod                  264
     nmap                 231
     guess_passwd          53
     buffer_overflow       30
     land                  21
     warezmaster           20
     imap                  12
     rootkit               10
     loadmodule             9
     ftp_write              8
     multihop               7
     phf                    4
     perl                   3
     spy                    2
     Name: label, dtype: int64
```

Result:

- The different types of attacks in the dataset
- This much number of counts the attack is happened
- In label mentioned that "normal" or "attack".



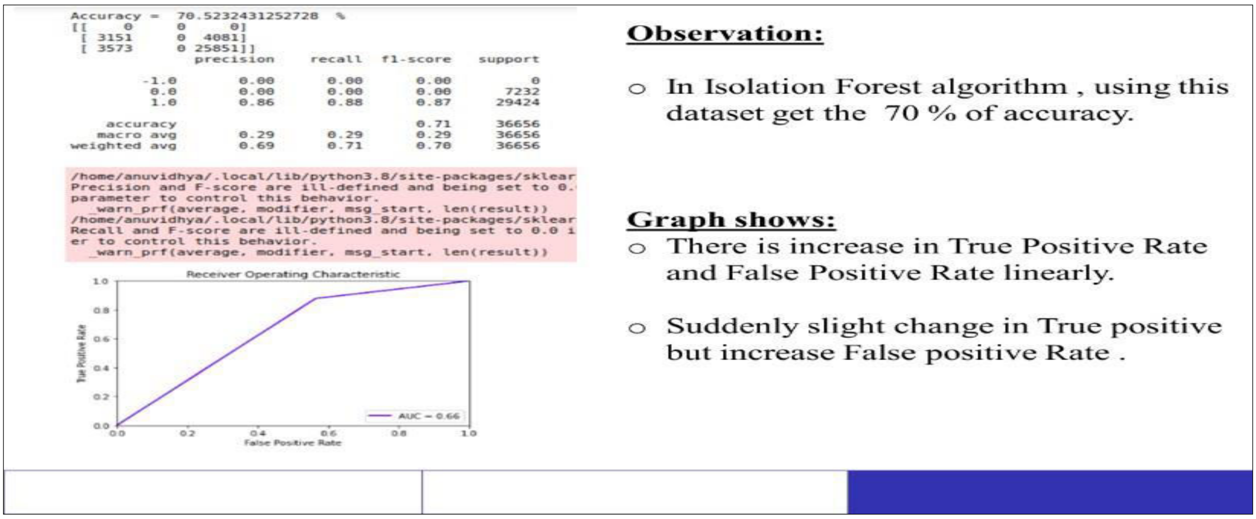
```
------------------------------------------------
Accuracy =  76.44587516368398  %
[[    0     0     0]
 [  492     0  6740]
 [ 1402     0 28022]]
              precision    recall  f1-score   support

        -1.0       0.00      0.00      0.00         0
         0.0       0.00      0.00      0.00      7232
         1.0       0.81      0.95      0.87     29424

    accuracy                           0.76     36656
   macro avg       0.27      0.32      0.29     36656
weighted avg       0.65      0.76      0.70     36656

/home/anuvidhya/.local/lib/python3.8/site-packages/sklearn
Precision and F-score are ill-defined and being set to 0.0
parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/home/anuvidhya/.local/lib/python3.8/site-packages/sklearn
Recall and F-score are ill-defined and being set to 0.0 in
er to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Observation:

- In Local Outlier Factor algorithm, using this dataset get the 76 % of accuracy.

Graph shows:

- It shows there is linear increase in True Positive Rate and False Positive Rate.



```
Accuracy =  70.5232431252728  %
[[    0     0     0]
 [ 3151     0  4081]
 [ 3573     0 25851]]
              precision    recall  f1-score   support

        -1.0       0.00      0.00      0.00         0
         0.0       0.00      0.00      0.00      7232
         1.0       0.86      0.88      0.87     29424

    accuracy                           0.71     36656
   macro avg       0.29      0.29      0.29     36656
weighted avg       0.69      0.71      0.70     36656

/home/anuvidhya/.local/lib/python3.8/site-packages/sklear
Precision and F-score are ill-defined and being set to 0.
parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/home/anuvidhya/.local/lib/python3.8/site-packages/sklear
Recall and F-score are ill-defined and being set to 0.0 i
er to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Observation:

- In Isolation Forest algorithm , using this dataset get the 70 % of accuracy.

Graph shows:

- There is increase in True Positive Rate and False Positive Rate linearly.
- Suddenly slight change in True positive but increase False positive Rate .

VII. CONCLUSION

As the involvement of the systems over the internet increasing rapidly, the security concerns have also seen. The proposed approachdeals with the detection of intruders over the internet efficiently. The proposed algorithm has performed well as compared to the previously applied algorithms such as SVM, Naive Bayes, and Decision Tree. The detection rates and the false error rates can be improved at a great extent by the proposed approach. The dataset used here is the knowledge discovery dataset.

REFERENCES

1. JafarAbo Nada; Mohammad Rasmi Al-Mosa, 2018 International Arab Conference on Information Technology (ACIT), A Proposed Wireless Intrusion Detection Prevention and Attack System.

2. Kinam Park; Youngrok Song; Yun-Gyung Cheong, 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigData Service), Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm.

3. S. Bernard, L. Heutte and S. Adam “On the Selection of Decision Trees in Random Forests” Proceedings of International JointConference on Neural Networks, Atlanta, Georgia, USA, June 14-19,2009, 978-1-4244-3553-1/09/$25.00.

4. A. Tesfahun, D. Lalitha Bhaskari, “Intrusion Detection using Random Forests Classifier with SMOTE and \eature Reduction”2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 978-0-4799-2235-2/13 $26.00 © 2013 IEEE.

5. Le, T.-T.-H., Kang, H., & Kim, H. (2019). The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection. 2019 International Conference on Platform Technology and Service (PlatCon). Doi:10.1109/platcon.2019.8668960.

6. Anish Halimaa A, Dr K.Sundarakantham: Proceedings of the ThirdInternational Conference on Trends in Electronics and Informatics (ICOEI 2019) 978-1-5386-9439- 8/19/$31.00 ©2019 IEEE “MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM.”

7. Mengmeng Ge, Xiping Fu, Naeem Syed, Zubair Baig, Gideon Teo,Antonio Robles- Kelly (2019). Deep Learning-Based Intrusion Detect ion for IoT Networks, 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 256-265, Japan.

8. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, “An Investigation on Intrusion Detection System Using Machine Learning” 978-1-5386-9276-9/18/$31.00 c2018IEEE.

9. Rohit Kumar Singh Gautam, Er. Amit Doegar; 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) “An Ensemble Approach for Intrusion Detect ion System Using Machine Learning Algorithms.”

10. Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahma, 2019 International Conference on Robot ics, Electrical and Signal Processing Techniques (ICREST)“Network Intrusion Detect ion using Supervised Machine Learning Technique with Feature Selection.”

11. .L. Haripriya, M.A. Jabbar, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)” Role of Machine Learning in Intrusion Detection System: Review”

12. Nimmy Krishnan, A. Salim, 2018 International CET Conference on Control, Communication, and Computing (IC4) “ Machine Learning-Based Intrusion Detect ionfor Virtualized Infrastructures”

13. Mohammed Ishaque, Ladislav Hudec, 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) “Feature extract ion using Deep Learning for Intrusion Detection System.”

14. Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar, Rashmi Bhattad, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)“A Review of Machine Learning Methodologies for Network Intrusion Detection.”

15. Iftikhar Ahmad , Mohammad Basheri, Muhammad Javed Iqbal, Aneel Rahim, IEEE Access (Volume: 6) Page(s): 33789 – 33795 “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for IntrusionDetection.”

16. B. Riyaz, S. Ganapathy, 2018 International Conference on Recent Trends in Advanced Computing (ICRTAC)” An Intelligent Fuzzy Rule-based Feature Select ion forEffective Intrusion Detectio n.”

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING