CHARLOTTE

THE GRADUATE SCHOOL

Anomali Detetcion in Cybersecurity

Applied ML project Description-Ashlesha Gupta

CHARLOTTE
THE GRADUATE SCHOOL

By - Ashlesha Gupta

- **Name** : Ashlesha Gupta
- **Term** : Fall 22
- **University** : University Of North Carolina At Charlotte
- **Student Id** : 801281891

Paper 0: CyberAttack Detection Using Anomaly Detection Technique in machine learning

- **Title** : **Detecting Cyber Attacks Using Anomaly Detection with explanations and expert feedback**
- **Authors** :Md Amran Siddiqui,Jack W. Stokes,Christian Seifert,Evan Argyle,Robert McCann,Joshua Neil,Justin Carroll.
- **Year Published**: 1 May 2019
- **Conference Name** : ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- **Location**: Brighton, UK

CHARLOTTE
THE GRADUATE SCHOOL

Paper 0: Cyber Attack Detection Using Anomaly Detection Technique in machine learning

Short Summary:

Here along with the anomaly detection , along with testing the efficiency of various algorithms like random forest, knn, svm etc. we will also try to develop a human-in-the-loop security attack detector using an anomaly detection technique that can provide explanation about the detected anomaly and can improve its detection capabilities using feedback from security experts. This will help improve the accuracy of algorithm and reduce the number of low fidelity and false positives.

- **Problem Statement**

Cybersecurity attacks are growing both in frequency and sophistication over the years.

Due to shortage of skilled professionals in cybersecurity it has become very difficult to investigate these attacks using traditional manual analysis methods .

This project will explore Machine Learning as a viable solution by examining its capabilities to classify malicious traffic in a network and help identify potential cyber threats and attacks .

Based on the final output of the result analyst can further classify the true-positive events which used to be a huge problem earlier as the analyst sometimes has no indications about why the particular computer was identified as being “under attack.

Data Sets That will be used in this project

- Netflow CTU-13 dataset:

  Link :https://mcfp.felk.cvut.cz/publicDatasets/CTU-13-Dataset

  About the dataset:

  The network traffic dataset known as CTU-13 was recorded in 2011 at the CTU University in the Czech Republic. The dataset's objective was to have a significant amount of real botnet traffic captured with normal traffic including background traffic.

- To cross- check the accuracy of the model we will also use BETH dataset:

  Link :https://www.kaggle.com/datasets/katehighnam/beth-dataset

  About the dataset:

  This data has DNS network traffic logs. It contains real-world attacks in the presence of benign modern OS and cloud provider traffic, without the added complexity of noisy artificial user activity .This data set is mainly used in anomaly detection algorithms

Data Sets That will be used in this project

- **UNSW-NB15 Dataset**

  Dataset link : https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15

  About the dataset:Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms are among the nine attack categories in this dataset. To produce a total of 49 characteristics with the class label, the Argus and Bro-IDS tools are utilized, and twelve methods are built.

Motivation .

Having worked in the cyber security industry as a security analysts and then as a Security Orchestration Automation and Response Engineer , I have realised that it's very difficult for a person to analyse tons of logs from thousands of security controls with low fidelity data . Analysts are bombarded with this low fidelity data that they don't get time to deeply analyse and use their human intelligence in threat hunting.Manually verifying logs is a costly and time consuming investigation.And that's why the threats and anomalies go undetected and that finally lead to cyber attacks. So ,to avoid all this we can definitely leverage the power of machine learning and this is my attempt to solve the information overload problem of the cybersecurity analysts so that they can focus on the analysis that requires indispensable human intelligence.For the protection of next-generation cyber machine learning is one of the essential component for activating enhanced degrees of cyber security.This will greatly help the Security Operations Team to focus more on what matters the most instead of drowning in the low fidelity data flood.

Here's to all the Security Engineers!

UNIVERSITY OF NORTH CAROLINA CHARLOTTE | THE GRADUATE SCHOOL

Paper 1 : Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network

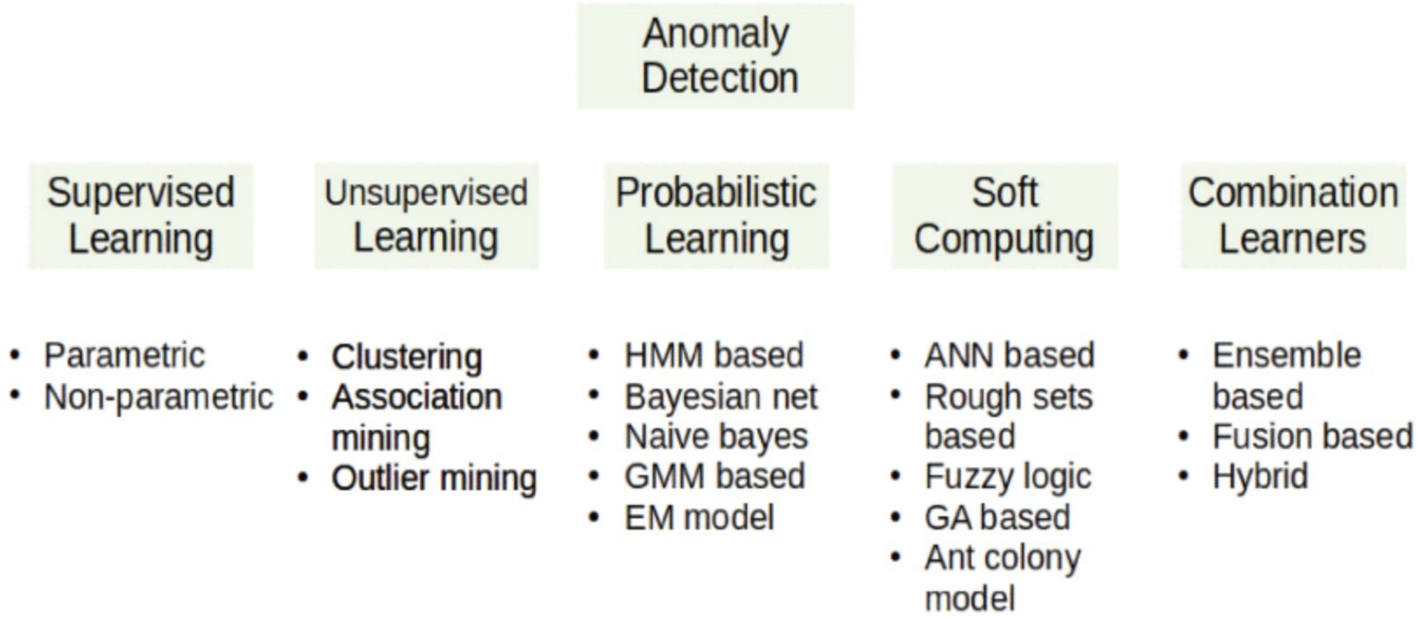
Citation: Tufan, Emrah & Tezcan, Cihangir & Acarturk, Cengiz. (2021). Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network. IEEE Access. 9. 50078-50092. 10.1109/ACCESS.2021.3068961.

Date Published : 26 March 2021

Location : IEEE Access conference.

Authors:Emrah Tufan,Cihangir Tezcan,Cengiz Acartürk

Paper 1 : Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network



Paper 1 : Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network

About The Paper:

This paper talks about the conventional rule based intrusion detection system and how they are fast and convenient but conventional IDS mechanisms are limited in their flexibility, thus being disadvantageous in detecting novel types of attacks.

They developed two ML models and trained the models by the data obtained from an institutional network dataset and UNSW-NB15 data set and various algorithms are implemented to test the efficiency of the algorithms.

Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.

Probabilistic learning models estimate new instances affected by randomness and other probabilistic uncertainty [9]. A probabilistic model’s distinctive feature is updating the previous estimates based on new evidence learned from training data. Probabilistic learning has been used for IDS design.

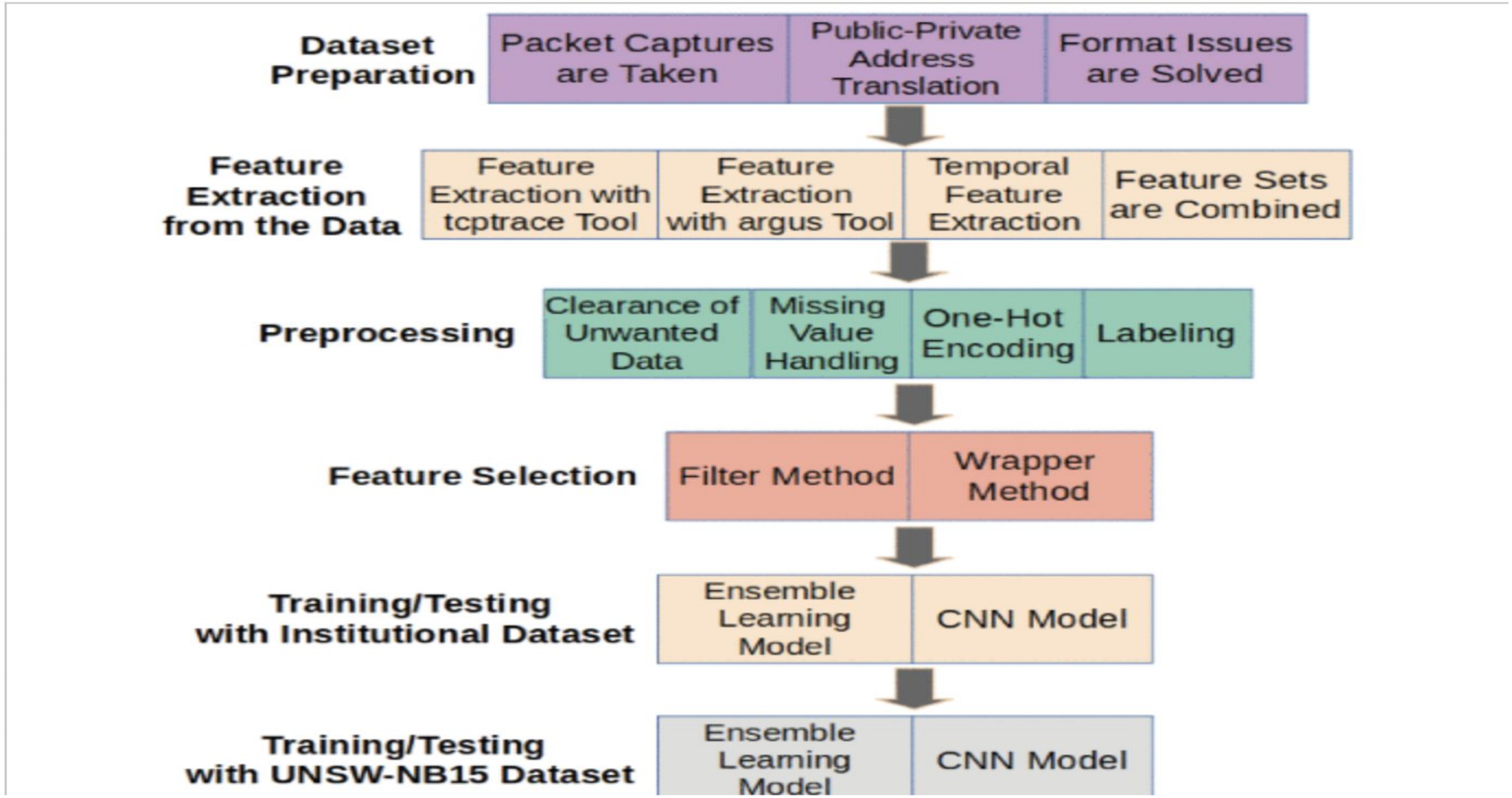
Paper 1 : Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network

After comparing the results of all the algorithms like random forest , logistic regression, KNN , SVM and CNN , the best was CNN in terms of efficiency and accuracy but some ML methods provide a more explainable structure about their inner mechanism while others do not.The former category is called “white box,” whereas the latter is “black box.”

Methods like decision trees, logistic regression, KNN, and SVM are positioned closer to the white-box since it is easy to deduce what the decision-making processes are like in these algorithms. Security is a crucial topic where less explicable ML models should be chosen with greater caution. And that’s why Convolutional Neural Networks is not a best choice . Another drawback of employing a CNN model is its computational and temporal costs.

The below diagram wonderfully illustrates all the steps taken in this research :

Paper 1 : Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network



UNIVERSITY OF NORTH CAROLINA CHARLOTTE | THE GRADUATE SCHOOL

Paper 1 : Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network

Why I chose this paper ? How exactly does the selected paper solve the problem?

I chose this paper because it gives a clear comparison of all the models (SVM,LR,KNN,CNN,Random Forest) etc. for evaluating an efficient Anomaly detection Algorithm and Intrusion Detection follows anomaly detection which I am going to work on. And it also compared the approaches followed in Rule based Intrusion Detection and Anomaly based intrusion detection. After comparing the efficiency of Models I was introduced to the new concept white box and black box methods for cybersecurity and this research paper will definitely complement the findings of my project .

The selected paper solves the problem by following the complete procedures like:

1.Data set Preparation.

2.Feature Extraction from the data

3.Preprocessing

4.Feature Selection

5.Training /Testing with both the data sets to get the efficiency and accuracy

Paper 2 : Network Based Intrusion Detection Using the UNSW-NB15 Dataset

Citation: al, Mefta. (2019). Network Based Intrusion Detection Using the UNSW-NB15 Dataset. International Journal of Computing and Digital Systems. 8. 477-487. 10.12785/ijcds/080505. .

Date Published : January 2021

Location : 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)

Gazipur, Bangladesh

Authors:Abu Saleh Md Towfiqur Rahman,Md. Mahbubur Rahman,Samrat Kumar Dey

Paper 2 : Network Based Intrusion Detection Using the UNSW-NB15 Dataset

Short Summary Of The Research:

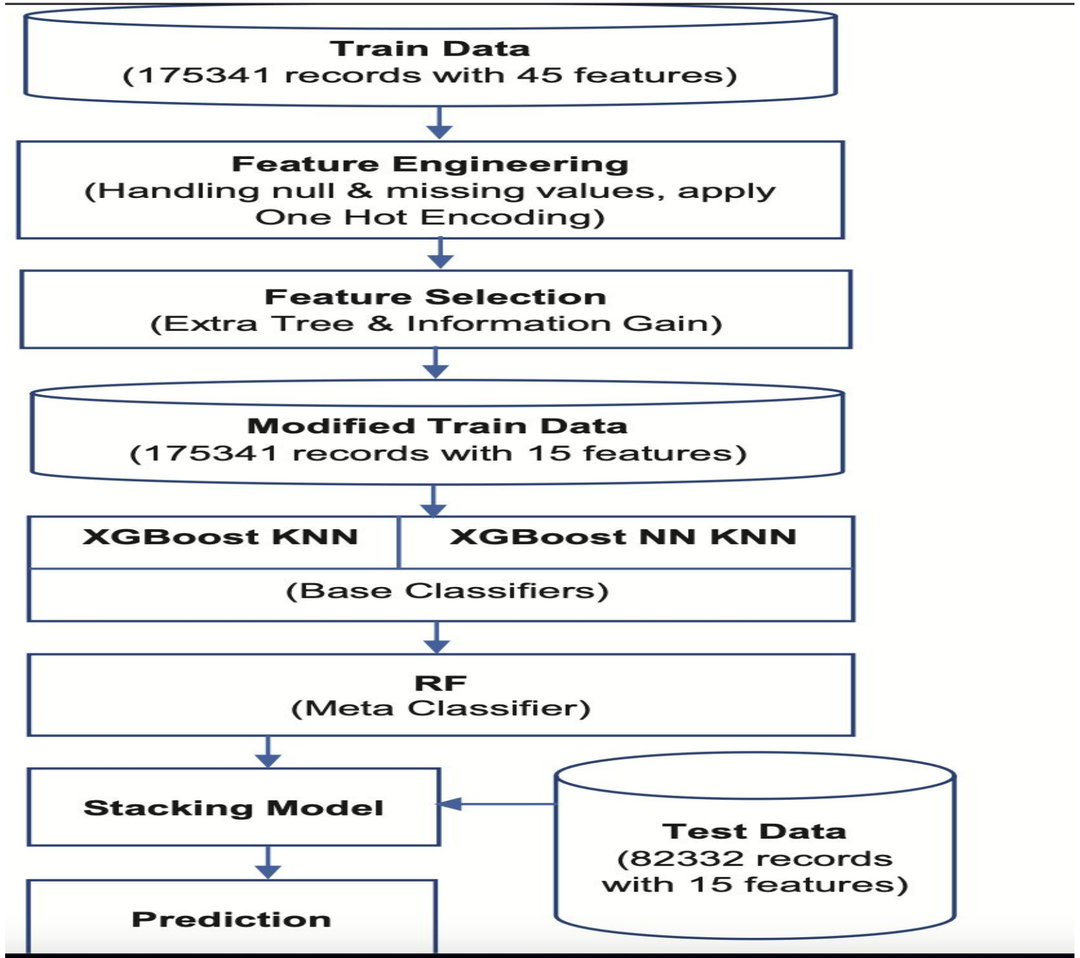
Signature based Network Intrusion Detection Systems(NIDS) can't detect zero day attacks as these attacks are not included in the database and signatures are also not in the database.But,an intrusion detection system that is capable of identifying anomalies uses the ML technique to recognize zero-day exploits .

This paper has implemented models on the UNSW-NB15 dataset utilizing complete training and testing data subset to train and test our models and built a NIDS with 96.24% accuracy .This paper has proposed two stacking ML models with feature selection (FS) .

Depending on the nature of ML algorithms, the researchers combined two ensemble algorithms (XGBoost and RF) with a simple supervised algorithm (KNN) and found that the stacking outperforms the individual models and also provides better accuracy .

Paper 2 : Network Based Intrusion Detection Using the UNSW-NB15 Dataset

Short Summary Of The Research: The following figure helps illustrate all the steps taken :



Paper 2 : Network Based Intrusion Detection Using the UNSW-NB15 Dataset

Short Summary Of The Research: The following figure helps illustrate all the steps taken :

In this research , two stacking models are developed by combining ensemble algorithms with simple supervised algorithms KNN and NN respectively. We have used the UNSW-NB15 dataset which has 175, 341 train and 82,332 test data. In this paper, a network intrusion detection system built with the stacking ML model is proposed where XGBoost and KNN act as base classifiers and RF as a meta classifier. It is experimentally proved that the proposed stacking model built with the combination of ensemble algorithms (XGBoost and RF) and relatively simple algorithm (KNN) outperforms all other recent competing models. Again, as our model is using fewer machine learning algorithms, it is likely to take less time to execute.

Paper 2 : Network Based Intrusion Detection Using the UNSW-NB15 Dataset

Short Summary Of The Research: The following figure helps illustrate all the steps taken :

Why I chose this paper ? How exactly does the selected paper solve the problem?

I chose this paper because it's employing more than 2 algorithms to solve the anomaly detection problem to increase the accuracy and efficiency of the final algorithm. This research paper will be really helpful while building my model from scratch as I am also going to use a combination of models to increase the accuracy and efficiency.

So , this paper is solving the accuracy and efficiency issue by stacking model built with the combination of ensemble algorithms (XGBoost and RF) and relatively simple algorithm (KNN) outperforms all other recent competing models.

UNIVERSITY OF NORTH CAROLINA CHARLOTTE | THE GRADUATE SCHOOL

TIMELINES

Tasks	Dates	Comments
Research Paper and Topic Selection and slides creation	12/09/2022 - 17/09/2022	Read various research papers to match the topic I selected for project and created short summaries of 2 of them based on my reading
Data set preparation	18/09/2022 - 02/10/2022	Data set preparation may require more time as the datasets can be raw and not
Feature Extraction From The data	03/10/2022 - 17/10/2022	Feature extraction is an important step to enhance the efficiency and accuracy of the model
Preprocessing The Data	18/10/2022 -05/11/2022	Preprocessing is another great step to normalise and regularize the data
Feature Selection	06/11/2022 - 19/11/2022	Feature selection may require more time
Training and Testing the Data with different Data Sets	20/11/2022 - 29/112022	Final step of the project