# NETWORK INTRUSION DETECTION USING

# MACHINE LEARNING ALGORITHMS

# A PROJECT REPORT

*for*

# INFORMATION SECURITY ANALYSIS AND AUDIT

# (CSE3501)

*in*

# B. Tech (Information Technology)

*by*

# KSHITIJ DHYANI (18BIT0131)

# Fall semester, 2020

*Under the Guidance of*

# Prof. SUMAIYA THASEEN I

Associate Professor Grade 1, SITE



# School of Information Technology and Engineering

* **LITERATURE SURVEY:**

1. **An Extremely Randomized Trees-Based Scheme for Stealthy Cyber-Attack Detection in Smart Grid Networks**

**Author:** MARIO R. CAMANA, SAEED AHMED 1,2, CARLA E. GARCIA, AND INSOO KOO

**Year:** 2019-2020

1. **Technique/algorithm used and why it was chosen (motivation)**

Smart grids are susceptible to cyber-attacks, and this could result in a complete monopoly of the system. The hacker could cheat a bad-detector and hence cause false control decision. To battle this issue the paper proposes a new and innovated approach based on extremely randomized trees algorithm.

1. **Architecture/ model/pseudocode developed**

A smart grid has countless sensors working in harmony to produce power, everything is kept in check using a Bad Data detector (BDD) which on encountering bad values (possible an attack) resets the system and protects it from incurring a loss. However, this generally requires the development of a very complex mathematical models. Hence in the recent times Machine Learning (ML) based detection tools are starting to gain a lot of traction, since they do not need any such complex models, they are generally fed the data of all the sensors of a grid in a steady, free and secure state, and the algorithm learns from it. In case any stealthy attack, the ML algorithm automatically detects a diversion from the ideal values and hence alerts the grid to reset.

1. **Datasets analyzed in the paper with the performance results**

The data set used for the study was SE-MF data set which corresponded to the meter measurements at a specific time.

**Performance results:**

For GBM, across four datasets:

Avg accuracy 118-Bus – 98.3%

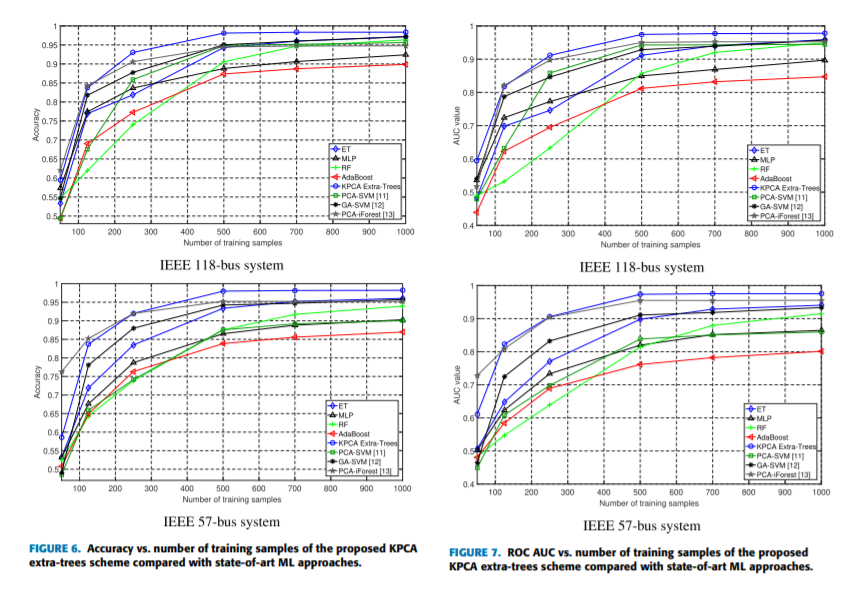
Avg AUC Value 118-Bus – 97.6%

Avg Accuracy 57-Bus – 98.15%

Avg AUC 57-Bus – 97.5%

1. **Any comparison done with the previous techniques to specify that the proposed method is superior**

The performance was compared to MLP, RF, AdaBoost, KPCA Extra Trees, PCS-SVM,GA-SVM, PCA-iForest algorithms and it was found that it out performs all other algorithms by a margin of quite significance.



1. **Android Malware Permission-Based Multi-Class Classification Using Extremely Randomized Trees.**

**Author:** Fahad Alswaina, Khaled Elleithy

**Year:** 2018-2019

1. **Technique/algorithm used and why it was chosen (motivation)**

Mobile phones have become an extension to the human body. This has resulted in an increased screen time and also regularly involves the usage of sensitive data of the users. Therefore, security and privacy has become increasingly important. The most basic and first line of defense against an intrusion is the android hardware and software permissions. The paper utilizes extremely randomized trees to create a reduced set of permission. This achieved a shorter execution time as well as a much faster execution.

1. **Architecture/ model/pseudocode developed**

The first and the most important step to block malicious software is to being successfully able to classify it as malicious software. The best step forward in this direction is to use Machine learning classification Algorithms, one such ensemble classifier called extremely randomized trees. Also the model inculcates the optimized nature by reducing the permission set by 18%, by classifying permissions based on their importance and likelihood to cause damage in case of intrusion. This enables the mobile application to only give the most required and essential permission, which in result reduces the chances of unnecessarily exposed vulnerabilities.

1. **Datasets analyzed in the paper with the performance results**

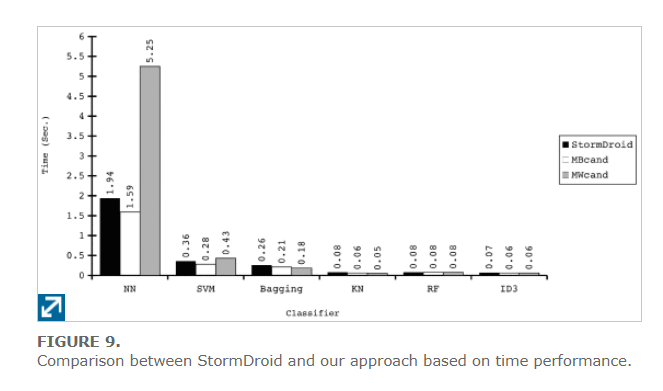
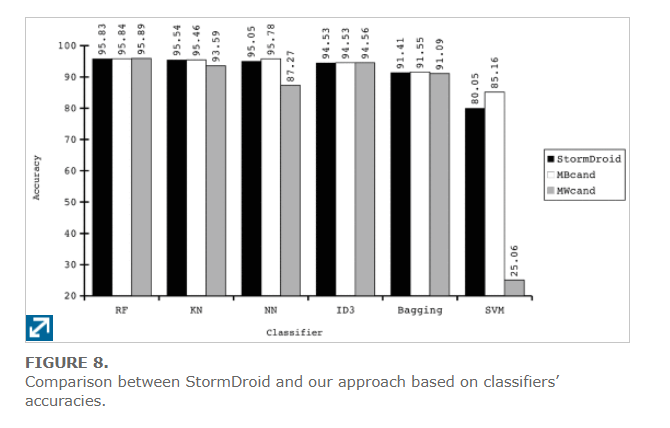
A customized data set was used in the study, it contained the general information about the families list, SFCand (a candidate subset of selected features SF ), MBcand (a two-dimensional binary matrix result from applying SFCand ), MWcand (a two dimensional weighted matrix result from applying the weight of each features in SFCand ), and *NoOfThreads* (number of threads set for framework efficiency; the default is 4)

**Performance results:**

Accuracy – 95.68 with standard deviation 0.19%

Precision- 95.99 %

1. **Any comparison done with the previous techniques to specify that the proposed method is superior**

The algorithm was compared to SVM, Bagging, ID3 , Neural Networks, KN against two variables which is accuracy and speed. The Random Forest based on randomized trees stood out to be the best in both the measures. 

1. **Ant Colony Induced Decision Trees for Intrusion Detection.**

**Author:** Frans Hendrik Botes, Louise Leenen, and Retha De La Harpe

**Year:** 2019-2020

1. **Technique/algorithm used and why it was chosen (motivation)**

As per the Moore’s law of exponential growth, data has been flowing through is massive amounts around the globe. We are a dynamic generation where things are keep growing, and so do the hackers. However, the field of Intrusion detection, security and privacy have been rather stagnant. Machine learning has great potential in the field of intrusion detection. This paper keeps the focus on the classifier based on decision trees using ant colony optimization instead of traditional CART techniques.

1. **Architecture/ model/pseudocode developed**

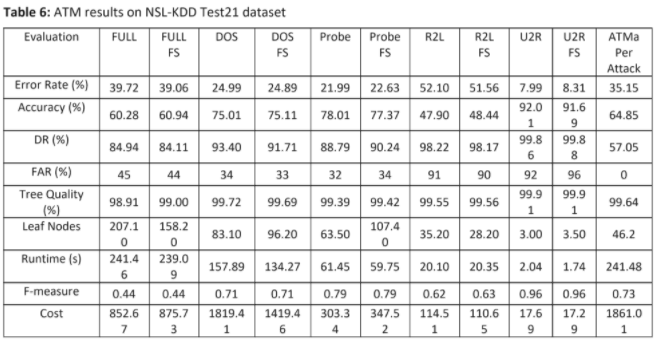
The real crux of cyber security is detection, and detection is in direct correlation with classification, since without proper classification, detection is quite difficult. Machine learning focuses on drawing massive amounts of data from a network. Then its tries to find patterns in the data set. In case of any anomality in the future the pattern is checked for reference, and its often disrupted and hence detection becomes a cake walk. The divide and conquer approach are used to induce a decision tree, however this paper uses and ATM classifier which results in a greater accuracy and precision.

1. **Datasets analyzed in the paper with the performance results**

The data set used is NSL-KDD dataset (Tavallaee et al. 2009) which is an improvisation upon the popular KDD99 dataset.

**Performance results:**

It has massive improvements in all measures.



1. **Any comparison done with the previous techniques to specify that the proposed method is superior**

In this paper, it was not compared to other algorithms. However it was tested against many attacks and proved to be an improvisation from the industry standards.

1. **Performance evaluation of intrusion detection based on machine learning using Apache Spark.**

**Author:** Arif Jamal Malik, Farrukh Aslam Khan

**Year:** 2017-2018

1. **Technique/algorithm used and why it was chosen (motivation)**

This paper focuses on comparing all the most famous intrusion detection methods that use machine learning algorithms such as SVMs, Naïve Bytes, Decision tree, and random forest. This paper gives us the insight on how tree-based algorithms have proven to be the best of the lot, and how randomized trees are up a notch higher. This paper uses apache spark to judge the algorithms on the basis of accuracy, sensitivity, specificity, training time, and as well as prediction time.

1. **Architecture/ model/ pseudocode developed**

The accuracy, sensitivity and specificity are calculated using four elements. Namely True positives (TP), true negative’s (TN), false positives (FP) and false negatives (FN).

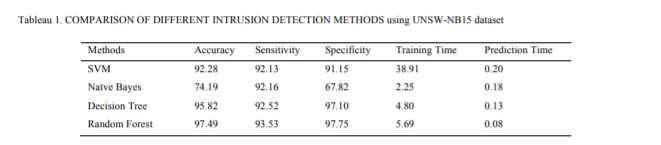


****

****

1. **Datasets analyzed in the paper with the performance results**

The paper also uses NSL-KDD dataset (Tavallaee et al. 2009) which is an improvisation upon the popular KDD99 dataset.



1. **Any comparison done with the previous techniques to specify that the proposed method is superior**

This paper is a comparative analysis between all the machine algorithms, does not compares any previous technique.

This study proves that random forest is overall best algorithm for intrusion detection when it comes to machine learning algorithms, since it has the highest accuracy, sensitivity as well as specificity. With a record training time and prediction time.

1. **A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection.**

**Author:** Deng, S., Wang, C., Wang, M., & Sun, Z.

**Year:** 2017-2018

1. **Technique/algorithm used and why it was chosen (motivation)**

Optimizing accuracy as well as the speed of detection and training have always been the major concern of all the machine learning algorithms. This paper focuses on pruning the decision tree, which is essentially reducing the size of the decision tree. This helps in reducing the complexity of the DT as well as the improves the accuracy.

1. **Architecture/ model/pseudocode developed**

The proposed technique is a hybrid approach in which PSO is used for node pruning and the pruned DT is used for classification of the network intrusions. Single layer as well as the multi-objective algorithms are used to make sure comprehensive study.

1. **Datasets analyzed in the paper with the performance results**

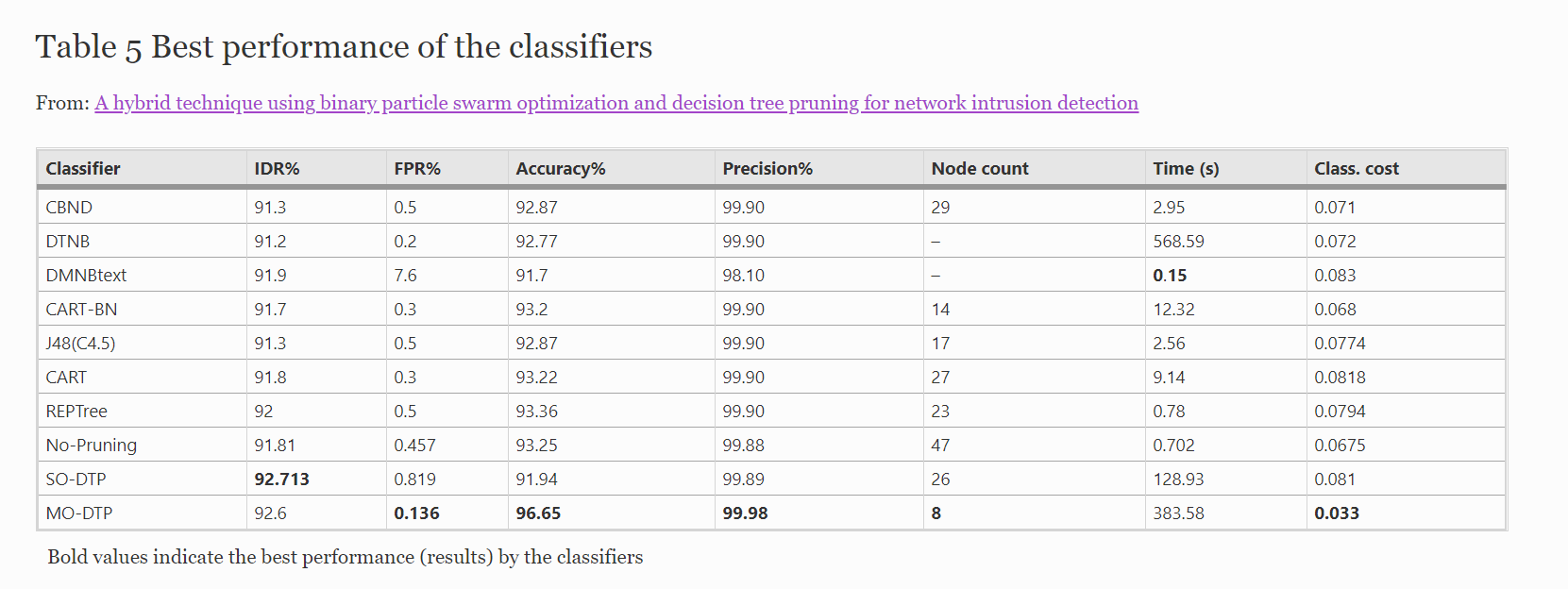
The data set used is a well-established one called KDD99Cup dataset. This dataset is an industry standard and very well respected by the cyber security community.

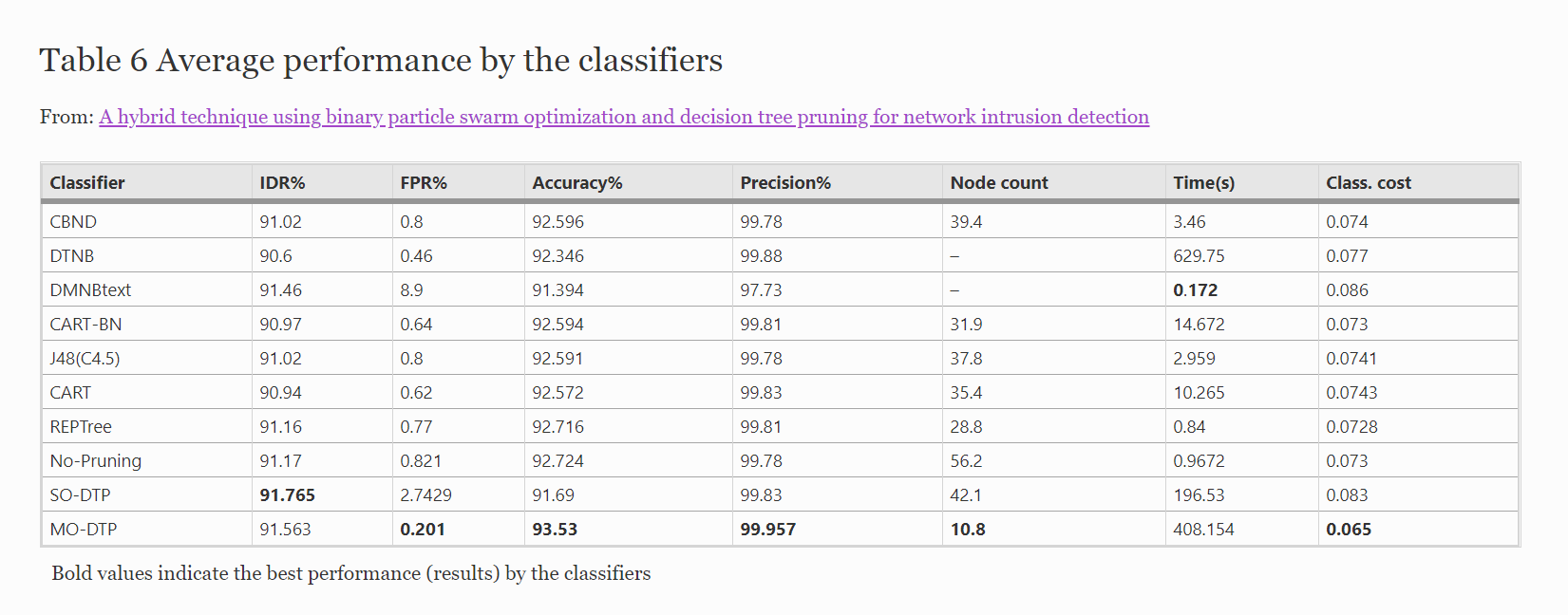
The results of the paper proved to be highly fruitful, indicating a increase detection rate, accuracy and precision. Also, the false positive rate dropped in number.

**Performance results:**

1. **Any comparison done with the previous techniques to specify that the proposed method is superior**

When compared to conventional DT techniques, a pruned DT showed better performance dude to various underlying reasons.





* **REFERENCES:**

[1] Acosta, M. R. C., Ahmed, S., Garcia, C. E., & Koo, I. (2020). Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks. *IEEE Access*, *8*, 19921-19933.

[2] Alswaina, F., & Elleithy, K. (2018). Android malware permission-based multi-class classification using extremely randomized trees. *IEEE Access*, *6*, 76217-76227.

[3] ECCWS 2017 16Th European Conference On Cyber Warfare And Security". *Google Books*, 2020, <https://books.google.co.in/books?hl=en&lr=&id=uFA8DwAAQBAJ&oi=fnd&pg=PA53&dq=Intrusion+detection+decision+trees&ots=YSw1lDVq-F&sig=xJxtoIRIm4AB-Zo1A0Iy-r-VQcA&redir_esc=y#v=onepage&q=Intrusion%20detection%20decision%20trees&f=false>.

[4] Belouch, M., El Hadaj, S., & Idhammad, M. (2018). Performance evaluation of intrusion detection based on machine learning using Apache Spark. *Procedia Computer Science*, *127*, 1-6.

[5] Malik, A.J., Khan, F.A. A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection. *Cluster Comput* **21,**667–680 (2018). <https://doi.org/10.1007/s10586-017-0971-8>