*International Journal of Engineering Applied Sciences and Technology, 2019*
**Vol. 4, Issue 6, ISSN No. 2455-2143, Pages 209-213**
Published Online October 2019 in IJEAST (http://www.ijeast.com)

# CALIBRATION OF VARIOUS OPTIMIZED MACHINE LEARNING CLASSIFIERS IN NETWORK INTRUSION DETECTION SYSTEM ON THE REALISTIC CYBER DATASET CSE-CIC-IDS2018 USING CLOUD COMPUTING

V. Kanimozhi,
Department of Computer Science,
Sathyabama Institute of Science and Technology,
Chennai, Tamilnadu, India

Dr. T. Prem Jacob,
Department of Computer Science,
Sathyabama Institute of Science and Technology,
Chennai, Tamilnadu, India

*Abstract*--- **Our paramount task is to examine and detect network attacks that are one of the daunting tasks because the variety of attacks are day by day existing in colossal number. The proposed system identifies the botnet attacks using the latest cyber dataset CSE-CIC-IDS2018 which is released by Canadian Establishment for Cybersecurity (CIC). The cyber dataset can be accessed on AWS (Amazon Web Services). The Cybersecurity datasets by CIC is world-wide well known. The realistic network dataset consists of all the modern and existing attacks such as Brute-force attacks and password cracking, Heartbleed, Botnet, DoS (Denial of Service), DDoS also known as Distributed Denial of Service, Web attacks i.e. vulnerable web app attacks, and infiltration of the network from inside. The objective of the proposed research is to identify one class classification of Botnet attacks. Botnet attack is a Trojan Horse malware attack which poses a serious security threat to the banking and financial sectors. Since a specific classifier could possibly work for such datasets so it is crucial to finish a comparative examination of classifiers in order to achieve the most noteworthy execution in such basic detection of network attacks. The proposed framework is to incorporate different classifier methods such as KNearset Neighbor classifier, Naïve Bayes, Adaboost with Decision Tree, Support Vector Machine classifier, Random Forest classifier, and Artificial Intelligence to distinguish a portrayal of botnet attacks on the recent cyber dataset CSE-CIC-IDS2018. Classifier results are provided as accurate precision of different classifiers. And furthermore, the proposed framework uses the Calibration curve is a standard approach in analytical methods which generates reliability diagrams to check the predicted probabilities of various classifiers are well calibrated or not. Finally, the displayed graph proves how well the artificial intelligence technique outperforms all the other classifiers. which generates reliability diagrams to check the predicted probabilities of various classifiers are well calibrated or not. Finally, the displayed graph proves how well the artificial intelligence technique outperforms all the other classifiers.**

*Keywords*-- **AWS, botnet, Calibration curve, CSE-CIC-IDS2018, various machine learning classifiers**.

## I. INTRODUCTION

Intrusion detection systems (IDSs) are accessible in several sorts; the 2 main types are the Host intrusion detection systems (HIDS) and network intrusion detection systems (NIDS). HIDS and NIDS are strategies for security the board for PCs and networks. In HIDS, hostile to danger applications, for example, firewalls, antivirus programming, and spyware- discovery programs are introduced on each system PC that has two-route access to the outside condition, for example, the Internet. In NIDS, hostile to risk programming is introduced just at explicit focuses, for example, servers that interface between the outside environment and the network portion to be secured.

All IDSs use 2 modes of operation Signature-based IDS and Anomaly-based IDS. The signature-based technique appearance at checksums and message authentication. Signature-based detection strategies are applied even as well by NIDS as by HIDS. A HIDS can verify log and config files for any surprising rewrites, whereas a NIDS can verify the packet checksums. Anomaly-based IDS detects appearance for surprising or uncommon patterns of activities. This class may be enforced by each host and network-based intrusion detection systems.

Malicious activity like denial-of-service attacks, port scans, and attacks by observance the network traffic are observed by NIDS. All arriving packets are read by NIDS and searches for any suspicious patterns. once threats area unit discovered, supported its severity, the

*International Journal of Engineering Applied Sciences and Technology, 2019*
**Vol. 4, Issue 6, ISSN No. 2455-2143, Pages 209-213**
Published Online October 2019 in IJEAST (http://www.ijeast.com)

system will take action like notifying directors. To watch and analyze network traffic to safeguard a system from network-based threats, NIDS has been employed. The proposed system is to detect botnet attacks which pose a serious threat to financial sectors by incorporating various classifier models on CSE-CIC- IDS2018 with better prediction.

## II. CSE-CIC-IDS2018 ON AWS

This is the latest and realistic cyber dataset by Canadian Establishment for Cybersecurity (CIC) in 2018. For intrusion detection and malware anticipation, datasets by CIC and ISCX have been utilized worldwide. The primary target of this dataset is to build up an orderly way to deal with produce different and far-reaching benchmark dataset for intrusion detection on the formation of client profiles which contain theoretical portrayals of occasions and practices seen on the system.

This dataset consists of seven distinctive attack situations: Botnet, Heartbleed, Brute-force, Denial of Service and Distributed Denial of Service, inside network infiltration and Web attacks. The assaulting framework incorporates 50 machines and the victim association has 5 divisions and thirty servers and 420 machines are incorporated. The dataset incorporates the catches organize traffic and framework logs of each machine, alongside 80 highlights removed from the caught traffic utilizing CICFlowMeter-V3 [1].

*Availability of Dataset on the website: http://www.unb.ca/cic/datasets/ids-2018.html*

## III. IMPLEMENTATION

### *A. Making an Artificial Neural Network with Anaconda, Jupyter Notebook and SciKit-Learn*

To fabricate this Artificial Neural Network, we use Anaconda 3.0 and the most recent Scikit form 0.19.1 and Pandas version 0.23.1 in Jupyter Notebook. It very well managed through pip or Miniconda (Package Manager) [3].

### *B. Steps to be followed to incorporate the various machine learning classifiers*

i. Preprocess the dataset by clearing noisy, irrelevant and missing data
ii. Split the dataset into Training and Testing model
iii. Make the training model to learn and fit the various machine learning classifiers
iv. Calculate the accuracy score by using the Testing model

## IV. INCORPORATING VARIOUS MACHINE LEARNING CLASSIFIERS ON CSE-CIC-IDS2018

Classification is the way toward predicting the class of given data focuses. Every method embraces a learning calculation to recognize a model that best fits the relationship between the training data and the testing data [14]. Subsequently, a key target of the learning calculation is to construct a predictive model that precisely foresee the class labels of already unknown records. The classification method is a deliberate way to deal with grouping models from input data. For instance, Decision tree classifiers, rule- based classifiers, Neural systems, Support Vector Machines, and Naive Bayes classifiers are a diverse strategy to tackle a classification issue.

### *A. Naive Bayes Classifier*

It is a classification model dependent on Bayes' Theorem which assumes the fact that is independent among class variables. In straightforward terms, a Naïve (Strong) Bayes classifier accepts that the specific feature of a class is irrelevant to the other features of the class.

A probabilistic classifier is a classifier that can predict, given a perception of an input, a likelihood distribution over a lot of classes, as opposed to just yielding the most likely class belong to it. Alongside straightforwardness, Naive Bayes is known to beat even advanced other classification techniques.

### *B. Random Forest Classifier*

Random Forest Classifier could be a tree-based graph that involves building many trees (decision trees), then combining their output to enhance the generalization ability of the model. the methodology of mixing trees is understood as an ensemble method. Ensembling is nothing however a mix of weak learners (individual trees) to supply a robust learner. It can be used to solve the issue of both categorical data in classification and continuous data in regression problems.

### *C. K-Nearest Neighbors Classifier*

The k-nearest-neighbors is the supervised classification model, the whole training dataset is created as a model for KNN ie. also called Instance-based learning (data rows). At the point when a prediction is required for an unknown or new data, the kNN algorithm will look through the model (training set) dataset for the k-most matching data or instances. The most matched prediction cases are reported and returned as the prediction for the unknown or new data where "k" is the quantity of neighbors (instances) it checks.

### *D. Support Vector Machine Classifier*

Support Vector Machine in short known as SVM classifier can be used for both classification and regression.

*International Journal of Engineering Applied Sciences and Technology, 2019*
**Vol. 4, Issue 6, ISSN No. 2455-2143, Pages 209-213**
Published Online October 2019 in IJEAST (http://www.ijeast.com)


The aim of SVM is to build a hyperplane in multi-dimensional space which uniquely classifies the record instances. Hyperplanes are said to be decision barriers that help classify the record instances. Instances falling on both aspect of the hyperplane may be accustomed to another variety of classes. Also, the hyperplane's dimensions rely upon the quantity of features. If the dimension is two, then the hyperplane is appeared to be a line. If the dimension is three, then it becomes a two-dimensional hyperplane. It is hard to imagine when the dimensions exceed three.

Support vectors are statistics factors which are closer to the hyperplane and affect the location and direction of the hyperplane. The use of those support vectors, the margin of the hyperplane can be classified. In SVM we maximize the margin between the statistics factor and the hyperplane and thus how we built the SVM Classifier in our proposed system.

*E. Adaboost with Decision Tree Classifier*

A decision tree is a tree-like structure with nodes speaking to where we pick a property and make an inquiry; edges speak to the appropriate responses to the inquiry, and the leaves speak to the output or class name. A dataset has been dissected into tiny subsets by the decision tree while in the meantime a correlated decision tree is steadily created. A decision tree point has no less than two branches and a leaf node point addresses a course of action or decision. The most noteworthy decision node in a tree which looks at to the best pointer called root node. Decision trees can manage both numerical and categorical data.

It is the first extremely fruitful boosting calculation produced for classification. It is the best beginning stage for comprehension boosting. It very well may be utilized related to numerous different sorts of learning calculations to enhance execution. The output of the other learning calculations ('weak learners') is consolidated into a weighted total sum that displays as the last output of the helped classifier. The most suited and consequently model utilized with AdaBoost are decision trees with one level. Since these trees are so short and just contain one decision for classification, they are frequently referred to as decision stumps

*F. Artificial Neural Network Classifier*

A neural framework contains units (neurons), arranged in layers, which convert a data input into some output. Each unit takes a piece of information, applies an (as often as possible nonlinear) ability to it and a short time later passes the output on to the accompanying layer [12].

Generally, the frameworks are portrayed to be feed-forward: a unit reinforces its respect all of the units on the accompanying layer, yet there is no feedback to the past layer. Weightings are associated with the layers running from one unit then onto the following, and it is these weightings produced in the training phase to regulate a neural framework to handle the particular problem.

Table Ⅰ. Accuracy and Different Metric Scores of Various Machine Learning Classifiers on CSE-CIC-IDS2018

| CLASSIFIER MODELS | ACCURACY | PRECISION | RECALL | F1 | AUC |
|---|---|---|---|---|---|
| **ARTIFICIAL NEURAL NETWORK** | **0.9997** | **0.9996** | **1.0** | **0.9998** | **1.0** |
| **RANDOM FOREST CLASSIFIER** | **0.9983** | **0.9992** | **0.9988** | **0.9992** | **0.9** |
| **KN NEIGHBOR CLASSIFIER** | **0.9973** | **0.998** | **0.9988** | **0.9984** | **0.998** |
| **SVM CLASSIFIER** | **0.998** | **0.9** | **0.9988** | **0.9994** | **0.999** |
| **ADA BOOST CLASSIFIER** | **0.9996** | **0.9996** | **0.9988** | **0.9992** | **0.9988** |
| **NAÏVE BAYES CLASSIFIER** | **0.992** | **0.9929** | **0.9976** | **0.9953** | **0.981** |

*International Journal of Engineering Applied Sciences and Technology, 2019*
**Vol. 4, Issue 6, ISSN No. 2455-2143, Pages 209-213**
Published Online October 2019 in IJEAST (http://www.ijeast.com)

## V. CALIBRATION CURVE

Right here is a sklearn plot comparing the calibration curves for some of the famous type algorithms. The best possible method of measuring the performance of a classifier's probability prediction on a dataset is using the calibration curve which is also referred to as a standardized curve. The calibration curve is created as follows.

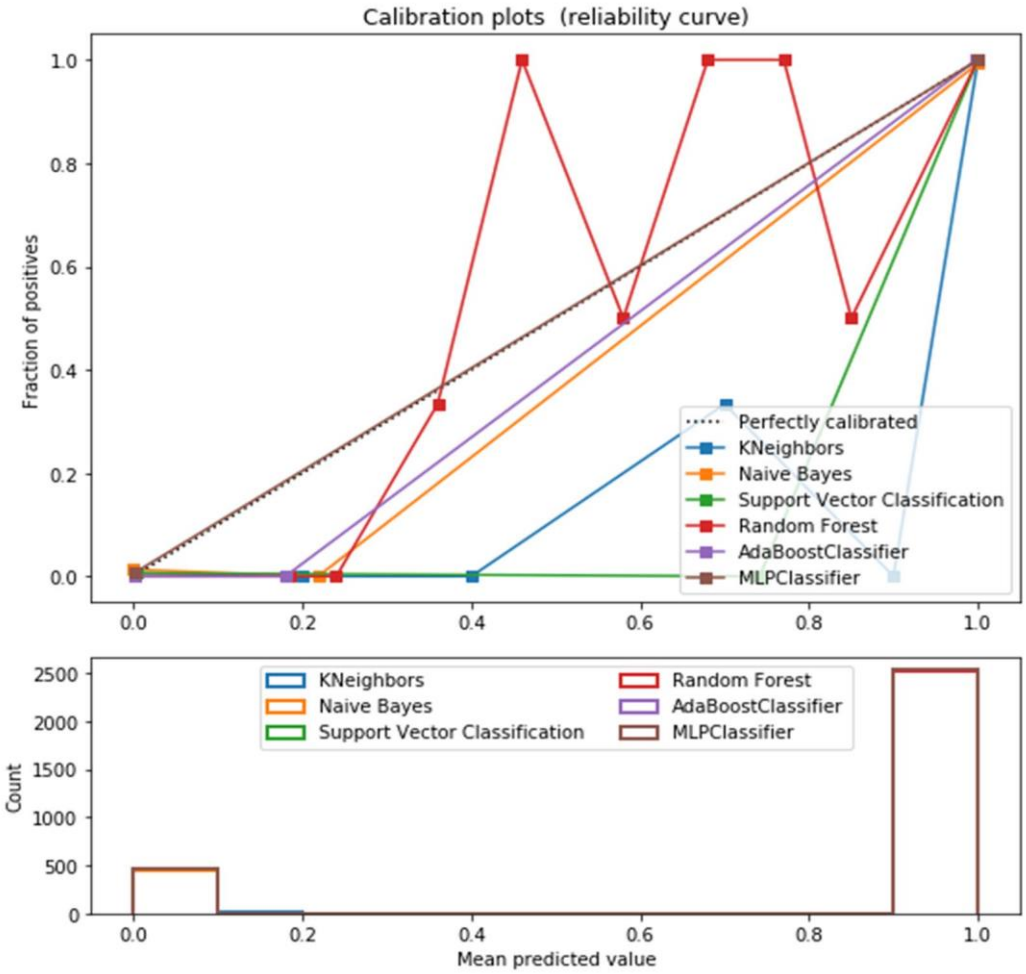


Fig 1. Calibration curve shows the MLP Classifier (Artificial Intelligence Technique) exactly spans on the dotted perfectly calibrated curve

The dotted line (standardized curve), shows the perfect calibrated curve. We have tested the various classifiers such as MLPClassifier, Naive Bayes, KNeighbors, Support Vector Classification, Random Forest, and AdaBoostClassifier for a model and tends to calibrate the best-performed classifier model on the cyber dataset. Calibration curves may be brought up as reliability diagrams.


The graph illustrates the deviation of six different classifier models from the perfectly calibrated curve [15]. It is difficult for logistic regression to handle large complex data. We use the dotted line to display the perfectly calibrated curve and the brown line denotes the MLP Classifier which exactly spans on the calibrated curve and is very close to the calibrated curve. And the curve which is second optimal fit is Adaboost classifier where the deviation is minimal and the third is Naive Bayes Classifier. SVM and KNeighbors are away from the standardized curve which shows maximal deviations. Even though SVM produces precision 1.0 after optimization but the Calibration curve produces the accurate performance of various classifiers. Random Forest classifier jumping from above to below and there is a strong tendency for samples to be slightly out of control and also shows the deviations with higher variance. The deviations in a graph are much easier to interpret at a glance. Thus Calibration curve is a huge data science mechanism for exploring the performance of various machine learning classifiers on this latest cyber dataset CSE-CIC-IDS2018.

## VI. CONCLUSION

The execution issue is a typical assignment when we go over pandas to work with bigger information (100 gigabytes to various terabytes), yet Spark is a publicly released Apache Framework utilized for huge information preparing can deal with parallel figuring with enormous datasets, running from 100 gigabytes to numerous terabytes crosswise over grouped PCs. The Research work can be further extended by classification of other significant existing attacks in this latest dataset.

## VII. REFERENCES

[1] Iman Sharafaldin, ArashHabibiLashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018.

[2] Alex Shenfield, David Day, and Aladdin Ayesh, "Intelligent intrusion detection system using artificial neural networks,"vol. 4, no.2, pp. 95-99, June 2018.

[3] Wu J., Peng D., Li Z., Zhao L., and Ling H. "Network intrusion detection based on a general regression neural network optimized by an improved artificial immune algorithm."Rev.,10 (3), 2015.

[4] Przemyslaw Kazienko & Piotr Dorosz. Intrusion Detection Systems (IDS) Part I - (network intrusions; attack ymptoms; IDS tasks; and IDS architecture). www.windowsecurity.com › Articles & Tutorials

[5] Sailesh Kumar, "Survey of Current Network Intrusion Detection Techniques", available at http://www.cse.wustl.edu/~jain/cse571-7/ftp/ids.pdf.

[6] Martin Roesch , "Snort – Lightweight Intrusion Detection for Networks", © 1999 by The USENIX Association.

[7] B. Daya ,"Network Security: History, Importance, and Future ,"University of Florida Department of Electrical and Computer Engineering , 2013.

[8] Liao H.-J., Lin C.-H.R., Lin Y.-C., and Tung K.-Y. "Intrusion detection system: A comprehensive review" Network Computing. Appl., Rev., 36 (1), pp. 16-24,2013. [Online].

[9] Antonia Nisioti, Alexios Mylonas, Paul D. Yoo, Vasilios Katos. "From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods",IEEE Communications Surveys & Tutorials, 2018 .

[10] Monowar H. Bhuyan, Dhruba K. Bhattacharyya, Jugal K. Kalita. "Network Traffic Anomaly Detection and Prevention", Springer Nature, 2017.

[11] JChristina Ting, Richard Field, Andrew Fisher, Travis Bauer."Compression Analytics for Classification and Anomaly Detection within Network Communication", IEEE Transactions on Information Forensicsa nd Security, 2018.

[12] Zhang G.P. "Neural networks for classification: A survey" IEEE Trans. Syst. Man Cybern. C, Rev., 30 (4), pp. 451-462, 2000.

[13] Singh R., Kumar H., Singla R.K., and Ketti R.R. "Internet attacks and intrusion detection system: A review of the literature"Online Inform. Rev., 41 (2), pp. 171-184, 2017.CrossRefView Record in ScopusGoogle Scholar

[14] Engen, Vegard. Machine learning for network based intrusion detection: an investigation into discrepancies in findings with the KDDCUP'99 data set and multi-objective evolution of neural network classifier ensembles from imbalanced data. Diss. Bournemouth University,2010

[15] D. Stiawan, A.H. Abdullah, and M.Y. Idris, "The trends of intrusion prevention system network, in: 2010" 2nd International Conference on Education Technology and Computer, vol. 4, pp.217-221, June 2010.

.