A Machine Learning Based Intrusion Detection Framework Using KDDCUP 99 Dataset

Diptee Agrawal[1], Prof. Chetan Agrawal[2], Prof. Himanshu Yadav[3]

[1]PG Scholar, [2,3]Assistant Professor

[1,2,3] Department of Computer Science & Engineering

[1,2,3] RITS, Bhopal, M.P., (India)

Abstract

The intrusion detection framework assumes a significant job in organization security. The Intrusion detection model is a prescient model used to foresee the organization's information traffic as regular or Attack. Machine Learning (ML) methods are utilized to assemble precise models for clustering, classification, and estimate. Intrusion detection systems (IDS) still have a few limitations and problems which opens new directions of research. As the era of the internet grows tremendously which results in more large data for which it needs to develop a better intrusion detection system that can work proactively. In this paper classification and prescient models for Intrusion detection are worked by utilizing ML based methods in particular BayesNet, NaiveBayes, J48, and Random Forest. These ML based methods are evaluated with KDDCUP 99 Dataset. Experiment results show that Random Forest Classifier beats different strategies in recognizing whether the information traffic is ordinary or an assault. We compute the results on various parameters like Accuracy, Precision, recall, F- Measures, and RoC area.

Keywords:- Machine Learning, Kddcup99, Intrusion detection system, Classification, Network Security.

Introduction

In now a time global market is ongoing toward the security of networks and systems. For that many big it giants make various intrusion detection systems for a high level of security, and various intrusion detection systems are presented in the market, but for every time attackers came up with the advanced method and breach the system integrity. More than the precedent 15 years, the increasing number of security incidents of a computer on the Internet has replicated the expansion of the Internet itself since the majority of deployed IT organizations are susceptible to assault. Cyber-attacks have been increasing exponentially with over 1.46 Billion recorded in 2019. The average injures from defenses violate was US$11.6m in 2017 as considered by NASDAQ.

Cyberspace is being turned up to computing appliances, networks, fiber-optics, wireless connection and other communications that convey the internet to billions of peoples around the world [4].those who use this technology they have various advantage over them in somehow the peoples who utilize the internet to tap into humanity’s collective store of knowledge every day but they don’t know about the darker side of the internet or information stealing. Data breaches are becoming ever bigger and more common. In 2018, over 1.4 billion records were lost, mainly through such attacks (see Fig. 1 for an exhibit of influenced industries) [4, 1].

Fig. 1 shows that all business areas that may affect by cyber-attacks. Individuals and multiple industries are two major industries that have been negatively impacted due to Cyber-attacks. Akamai released an innovative security testimony that grants scrutiny and insight into the worldwide assault hazard landscape. Akamai scrutinized a 52% boost in average peak bandwidth of assaults evaluated to Quarter 4 a year ago [4].

Detection of the intrusion is now a day's having big industry or study area also because of that we have to secure our information by attackers. In now a day's various cyber-attacks are being recruited and employed every day just because of predicting the various ways to breach system data security and also recognize new patterns to cheat systems in place for high-level security [5, 8]. This brings about a strong requirement of predicting such attacks proactively by recognizing those patterns using data analytics on historical data. Most of the Intrusion detection systems today generate a large number of alerts per second accumulating a large volume of continuous data that requires to be processed promptly [2, 6]. This explosion of data has led many researchers to discover hidden patterns and derive meaningful relationships within the data using data mining techniques. Many data mining and machine learning algorithms have been used to carry out this effort. The main emphasis is to detect as many attacks as possible with a minimum number of false alarms i.e., the system must be accurate in detecting the attacks [6, 3].

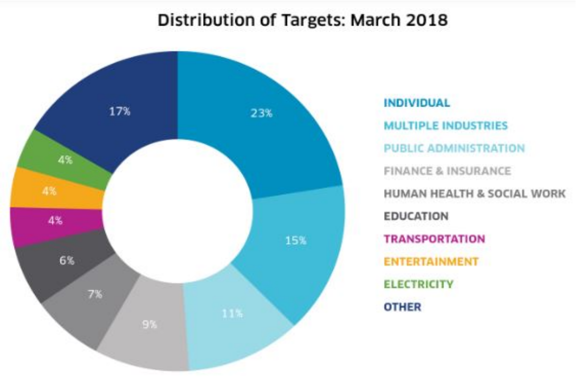


Fig. 1: Industries Targeted by Cyber attacks.

II. Literature Survey

Manjula C. Belavagi et al [7] express the Intrusion revelation model is a farsighted model used to envision the framework data traffic as common or Intrusion. Artificial intelligence counts are used to building exact models for gathering, portrayal, and desire. In their paper, request and judicious models for Intrusion acknowledgment are worked by using ML course of action estimations to be explicit Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, and Random Forest. These computations are attempted with the NSL-KDD instructive assortment. Their Experimental results show that Random Forest Classifier beats various procedures in perceiving whether the data traffic is average or an attack.

SMH Bamakan et. al. [11] states numerous associations perceive the need of using refined devices and frameworks to secure their PC arranges and diminish the danger of trading off their data. Albeit much ML-based information, grouping calculation has been proposed in-organize Intrusion recognition issue, every one of them has its qualities and shortcomings. Right now, propose a powerful Intrusion location system by utilizing another versatile, strong, exact improvement technique, in particular, time-changing disarray molecule swarm streamlining (TVCPSO) to all the while do parameter setting and highlight determination for various criteria straight programming (MCLP) and bolster vector machine (SVM). In the proposed techniques, a weighted target work is given, which considers exchange off between boosting the identification rate and limiting the bogus alert rate, alongside thinking about the number of highlights. Moreover, to make the molecule swarm improvement calculation quicker in looking through the ideal and dodge the pursuit being caught in nearby ideal, the tumultuous idea is embraced in PSO and time-differing dormancy weight and time-shifting increasing speed coefficient is presented. The exhibition of proposed strategies has been assessed by leading trials with the NSL-KDD dataset, which is gotten and altered from notable KDD cup 99 datasets.

The experimental outcomes show that the proposed technique performs better as far as having a high discovery rate and a low bogus alert rate when contrasted and the acquired outcomes utilizing all highlights.

Adel Sabry Eesa et. al. [12] presents another element choice methodology dependent on the cuttlefish improvement calculation which is utilized for Intrusion identification frameworks (IDSs). Since IDSs manage a lot of information, one of the essential undertakings of IDSs is to keep the best nature of highlights that speak to the entire information and expel the repetitive and immaterial highlights. The proposed model uses the cuttlefish calculation (CFA) as an inquiry procedure to find out the ideal subset of highlights and the choice tree (DT) classifier as a judgment on the chose highlights that are created by the CFA. The KDD Cup 99 dataset is utilized to assess the proposed model. Right now, researched the blended model of CFA and DT to include a choice for Intrusion location and assessed its presentation dependent on the benchmark KDD Cup 99 Intrusion information. Initially, we have changed the CFA to be utilized as a component choice instrument. At that point, they utilized DT classifier as an estimation of the produced highlights. Experimental outcomes uncover that the delivered highlights are played out the Detection Rate (DR) and Accuracy Rate (AR) (particularly when the quantity of created highlights was equivalent or under 20 highlights. By and large, at whatever point the quantity of highlights is diminished, the AR and DR are expanded. The outcomes show that the element subset got by utilizing CFA gives a higher location rate and exactness rate with a lower bogus caution rate when contrasted and they got outcomes utilizing all highlights.

Container Luo et. al. [9] present, a four-edge star-based imagined highlight age approach, FASVFG, is proposed to assess the separation between tests in a 5-class characterization issue. In light of the four-point star picture, numerical highlights are produced to arrange visit information from KDDcup99, and an effective Intrusion recognition framework with fewer highlights is proposed. The FASVFG-based classifier accomplishes a high speculation precision of 94.3555% in approval explore, and the normal Mathews connection coefficient arrives at 0.8858. Contrasted and the past IDSs, the key improvement of this new Intrusion recognition framework comes from the novel element age approach with the perception system. As we referenced previously, representation is an instinctive path for include determination and highlight decrease. The principal commitments of this work comprise of two sections.

First is the visual reform for high dimensional information. Through the proposed FASVFG framework, individuals are equipped for mapping an obscure system visit into a geometric point in the four edge star and deriving whether it is a typical visit or a specific malignant assault. The closer the fact is to one vertex, the more conceivable it could have a place with an assault classification. This methodology changes advanced system activity into a visual vision and offers an important and successful pre-information for a short time later classifier. Second is a significant endeavor to include a decrease. As appeared in test results, FASVFG is effectively tackling the 5-class characterization issue, in which the grouping exactness of the IDS accomplishes 94.3555%, and the normal MCC esteem accomplishes 0.8858. Even though the MCC estimation of FASVFG is somewhat lower than that with full highlights, to be specific, 0.9, the size proportion of highlight space is 16/43, which guarantees a progressively productive identification rate.

Salma Elhag et. al. [14] states Security strategies of data frameworks and systems are intended for keeping up the honesty of both the privacy and accessibility of the information for their confided clients. Be that as it may, various malevolent clients break down the vulnerabilities of these frameworks to increase the unapproved get to or to bargain the nature of administration.

Consequently, Intrusion Detection Systems have been structured to screen the framework and trigger alarms at whatever point they found a suspicious occasion.

Ideal Intrusion Detection Systems are those that accomplish a high assault recognition rate together with few bogus alerts. Notwithstanding digital assaults present a wide range of attributes that make them difficult to be appropriately recognized by straightforward factual techniques. As indicated by this reality, Data Mining strategies, and particularly those dependent on Computational Intelligence, have been utilized for actualizing vigorous and precision Intrusion Detection Systems.

Right now, consider the utilization of Genetic Fuzzy Systems inside a pairwise learning structure for the improvement of such a framework. The upsides of utilizing this methodology are twofold: first, the utilization of fluffy sets, and particularly etymological marks, empowers a smoother halfway point between the ideas and permits higher interpretability of the ruleset. Second, the separation and-overcome learning plan, in which we differentiate every single imaginable pair of classes with points, improves the accuracy for the uncommon assault occasions, as it acquires a superior distinguishableness between an "ordinary action" and the distinctive assault types. The integrity of our technique is bolstered by methods for a total test study, in which we differentiate the nature of our outcomes versus the best in a class of Genetic Fuzzy Systems for Intrusion identification and the C4.5 choice tree.

Ning Cao et al [10] state with the appearance of huge information time, information mining systems has been broadly used to assemble models for digital security applications, for example, spam separating, malware or infection discovery, and Intrusion recognition. This undertaking proposes a novel methodology that utilizes arbitrariness to improve the power of information mining models utilized in digital security applications against assaults that attempt to avoid discovery by adjusting. Their methodology tends to three issues. To start with, they manufacture a differing pool of mining models to improve the vigor of an assortment of mining calculations. These strategies are like outfit adapting however enhance the tradeoffs between mining quality and vigor. These strategies likewise require almost no adjustment to existing calculations. Second, they arbitrarily select a subset of models at run time (when the model is utilized for recognition) to additionally help heartiness. Third, they propose a hypothetical structure that limits the negligible number of highlights an assailant needs to alter given a lot of chosen models.

Yi Aung et al [13] states the security of the PC framework is vital, And over the most recent couple of years, there has seen an influenced development in the number of Intrusions that Intrusion identification has become the prevailing of current data security. Firewalls can't give total assurance. Applying on a firewall framework alone isn't sufficient to keep a corporate system from a wide range of system assaults. In this way, more frameworks ought to be supplemented by the Intrusion recognition framework. Information mining aptitudes can be utilized as a compelling way to deal with distinguishing Intrusions in an Intrusion discovery framework. Information Mining and Knowledge Discovery is the modernized procedure of digging and examination of tremendous measures of information, and afterward, extricate the importance of the information. Information mining apparatuses can help to anticipate future practices and patterns, with the goal that associations proactively, can settle on choices dependent on information. Information mining can address association addresses that were too conventional time, to comprehend. Information mining takes its name from the important data in a huge database. Late investigations show that falling based methodologies of a few calculations are vastly improved execution than an individual calculation. Right now, use K-means and Random Forest calculation to arrange occasions. This model was checked utilizing the KDD'99 informational collection. Test results show that crossbreed techniques can bolster appropriate identification rates and lower model preparing time than utilizing a solitary calculation.

III. Proposed Method & Tool

The underlying advance for the proposed explore system is choosing the KDD'99 dataset (Here, DARPA and NSL KDD-like dataset will likewise be considered for the examination) and afterward the instrument picked for reproduction is an information mining apparatus known as WEKA. The pre-processing of the chose (previously mentioned) dataset is finished with the assistance of shell programming of the previously mentioned technique. Further, five parameters are chosen for the observational outcomes. Under the recreation procedure, an arrangement model must be developed which is prepared on the preparation set. When the classifier model is prepared, it is prepared for reform on the undeveloped information.

Our Proposed work comprises of two fundamental modules, first dataset preparation utilizing the shell programming and the subsequent one is classification module, and the depiction of every module is shown beneath:

A) Dataset Preparation: The necessity for information pre-processing could be seen from the detail that pointless information and insignificant highlights may now and again puzzle the classification calculation, essentially to the location of incorrect or futile data. Besides, the preparation time will expand when each component is utilized. At long last, pre-handling helps to remove pointless information, insufficient information, and changes over the information into an institutionalized organization. The pre-preparing part of the proposed plot executes the accompanying functionalities:

- Executes repetition confirmation and handles every single invalid worth
- Translates downright information to numerical information

In the pre-preparing step, we convert the 10% KDD CUP 99 dataset which was in. CSV group into.ARFF position. By defaults KDD CUP 99 dataset is arranged of 5 kinds of assault that is Normal, DOS, R2L, U2R, and Prob, all things considered right now KDD CUP 99 dataset is pre-processing as a substitute of exaggerated 5 assault classes to 22 assault subcategories.

B) Classification: The yield of the previously mentioned calculation is gathered and named as a prepared dataset that has all assaults in a standard type of individual assaults. This dataset is currently prepared for the following procedure for example for the order. For this order, the WEKA 3.8 classifier component is utilized alongside the four distinct types of classifiers. Every single such analysis and prepared on 'WEKA 3.8' a broadly realized Machine Learning Workbench. Different prerequisites for the examinations were, the accessibility of Intel Core i7 Processors, 8 GB RAM, 1TB HDD, and UBUNTU 16 Operating System those are effectively fulfilled. All the classifiers were run with default parameters having 10 folds cross approval for model structure. The calculation given underneath clarifies the previously mentioned, bit by bit grouping process.

Flow- chart of proposed methodology can be shown below:

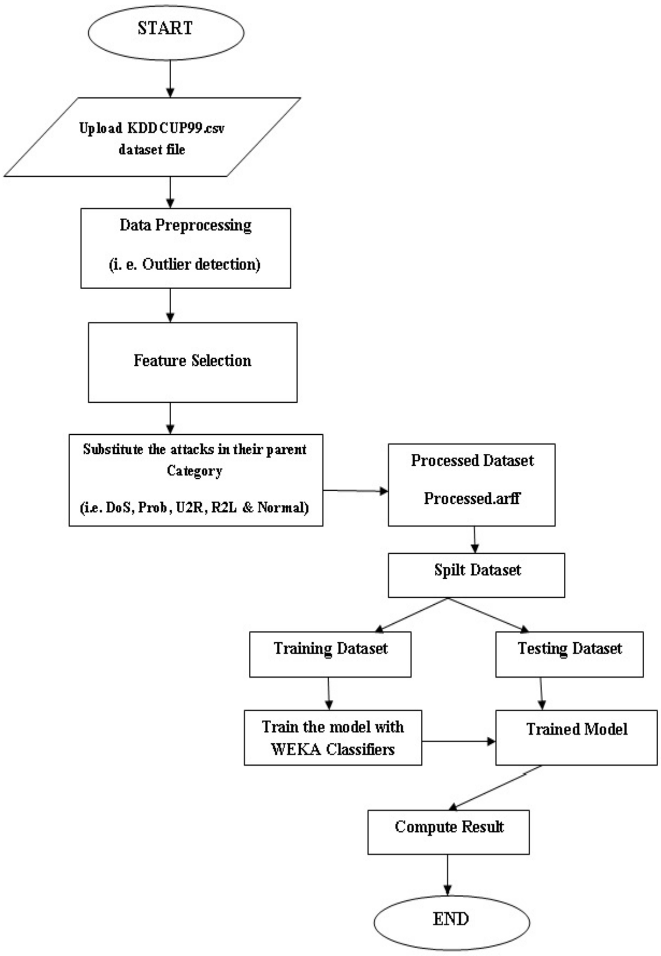

START
Upload KDDCUP99.csv dataset file
Data Preprocessing
(i. e. Outlier detection)
Feature Selection
Substitute the attacks in their parent Category
(i.e. DoS, Prob, U2R, R2L & Normal)
Processed Dataset
Processed.arff
Spilt Dataset
Training Dataset
Testing Dataset
Train the model with WEKA Classifiers
Trained Model
Compute Result
END

Fig.2: Flow Graph of Proposed Method.

There are several steps to pursue to apply the proposed method are as following:-

- Step 1- Start WEKA.
- Step 2- Upload the KDDCUP 99 dataset which is a combination of normal and abnormal dataset
- Step 3- Apply the Data preprocessing i.e. outlier detection etc.
- Step 4- do feature selection for appropriate categorization.
- Step 5- Replace the attacks with their parent Attack category.
- Step 6- produce a processed dataset file that has a parent category of intrusion attacks.
- Step 7- split the processed dataset into training and testing datasets (i.e. 70% for training and 30% for testing).
- Step 8- Apply WEKA classifiers one by one for classification (BayesNet, NaiveBayes, J48, Random Forest)
- Steps 9- we evaluate and compute all recital parameters with all WEKA classifiers and prefer the superlative best classifier.

The Classifiers utilized for exploratory investigation: The KDD Cup 99 dataset is relating to different WEKA classifiers, which we used for the experiment are

- BayesNet
- NaiveBayes
- J48
- Random Forests

BayesNet

The BayesNet model uses a directed graph in which particular edges show the relationships, from anyone can efficiently conduct inference on the random variables in the graph through the use of some factors. Before going into exactly what a Bayesian network is, it is first useful to review probability theory. Under this joint probability distribution of random variables A0, A1 …, An may be denoted as:

P(A0, A1, …, An) = P(A1 | A2, …, An) * P(A2 | A3, …, An) * … * P(An).

That is also called the chain rule of probability. And thus its factorized representation can be written as.

$$P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} P\left(A_k \,\middle|\, \bigcap_{j=1}^{k-1} A_j\right)$$

Moreover, the conditional independence between two random variables, A and B, given another random variable, C, is equivalent to satisfying the following property:

P (A, B|C) = P (A|C) * P (B|C).

Naive Bayes

Naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector x = (x1, …, xn) representing some n features (independent variables), it assigns to this instance probabilities for each of K possible outcomes or classes.

$$p(C_k|x_1,\ldots\ldots,x_n)$$

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We, therefore, reformulate the model to make it simpler. Using Bayes theorem, the conditional probability can be decomposed as −

$$p(C_k|x)=\frac{p(C_k)p(x|C_k)}{p(x)}$$

.

This means that under the above independence assumptions, the conditional distribution over the class variable C is −

$$p(C_k|x_1,\ldots\ldots,x_n)=\frac{1}{Z}p(C_k)\prod_{i=1}^{n}p(x_i|C_k)$$

J48

J48 is nothing but a famous type of decision tree algorithm which is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. J48 algorithm is used to create uni-variate decision trees to generate a decision tree. It uses a greedy technique to induce decision trees for 20 for reduced-error pruning. Moreover, the decision trees generated by J 48 can be used for classification, and that is why it is often referred to as a statistical classifier algorithm also. J48 algorithm is also inclined towards data mining techniques. This algorithm uses a greedy technique to induce decision trees for 20 classifications and uses reduced-error pruning.

Random forest

Random forest is a supervised learning algorithm that is used for both classifications as well as regression. However, it is mainly used for classification problems. Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution through voting. It is an ensemble method that is better than a single decision tree since it reduces the over-fitting by averaging the result. This model uses two key concepts that give it the name random:

1. A random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

A random sampling of training signifies that each tree here in the training process learns from a random sample of the data points. Where these samples are drawn with replacement, known as bootstrapping, and at test time, predictions are made by averaging the predictions of each decision tree. This procedure of training each learner on different bootstrapped subsets of the data and then averaging the predictions is known as bagging.

IV. Result Analysis

Investigation and trial perception here present the general results of our proposed calculation for the Intrusion identification framework. As recently referenced we executed our work in the Linux shell condition and from that point tried one of the most rumored datasets of Machine Learning (ML) known as KDD CUP 99. In our work, we have estimated numerous highlights and parameters identified with the classifiers and they are discovery accuracy, precision, recall, F-measure, and ROC for the various ML classifiers, individually. We have applied here, the directed procedure of ML by giving 10% of KDD Cup'99, for example, dataset having 321K occasions approx. as preparing inputs at that point, watching the anticipated ordinary and anomalous from the classifier mode as yields information, and from that point confusion matrix is framed by these anticipated results.

- Performance parameters:

True Positive (TP) / Recall: It is fundamentally the general extent of models that were delegated class x, among all models which genuinely have class x, for example, how many pieces of the class were caught. It is proportionate to Recall. In the disarray framework, this is the inclining component partitioned by the whole over the pertinent line,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

False Positive (FP): This rate is the genuine extent of models that were delegated class x, yet had a place with an alternate class, among all models which are not of class x. In the lattice structure, this is the segment total of class x less the corner to corner component, isolated by the aggregates of the columns of every single different class;

$$\text{FP} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

True Negative (TN): It alludes to the extent of negatives cases that were ordered effectively, this is determined by the accompanying condition:

$$\text{TN} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

False Negative (FN): It's only the extent of positive cases that were mistakenly delegated negative, and determined as follows:

$$\text{FN} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

Precision: - The Precision essentially the extent of the models which genuinely have class x among every one of those which were named class x. In the grid structure, it is spoken to the askew component isolated by the aggregate over the important segment,

$$\text{Accuracy} = \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

F- Measure: - one of the significant estimations in the investigation that joins accuracy and review is the consonant mean of exactness and review, the conventional F-measure or adjusted F-score. F-Measure that blends exactness and review is the symphonious mean of accuracy and review is known as F-measure.

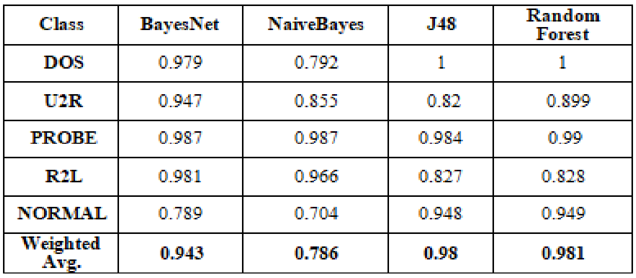
$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC: - - Receiver working attributes (ROC) diagrams help sort out classifiers and picture their presentation. Beneficiary Operating Characteristic (ROC), or ROC bend, is a graphical plot that outlines the exhibition of a twofold classifier framework as its segregation limit fluctuates. The bend is made by plotting the genuine positive rate against the bogus positive rate at different limit settings. Recipient Operator Characteristics (ROC) shows the exchange off of affectability and explicitness. ROC bends plot the genuine positive rate versus the bogus positive rate, at different edge shorts. The ROC is otherwise called a relative working trademark bend since it is an examination of two working qualities (TPR and FPR) as the model changes.

V. Experimental Results and Discussion

The Experiment Results investigation of the BayesNet, NaiveBayes, J48, and Random Forest classifiers is parted within Tables 1, 2, 3, 4, and 5. As could be seen the presentation of NaiveBayes Classifier is lower normal. For U2R and R2L assault is it's fewer than 43% score. The reason for this is because of the speculation of the NaiveBayes approach that all parameters are self-overseeing. All things considered, this isn't always the situation. Numerous assurance parameters are commonly reliant on each other. As a result NaiveBayes Classifier, even it takes a lesser measure of memory and is quicker in the figuring is maintained a strategic distance from under poor outcomes.

Table 1: Results of TP Rate of Each used WEKA Classifier



Class	BayesNet	NaiveBayes	J48	Random Forest
DOS	0.979	0.792	1	1
U2R	0.947	0.855	0.82	0.899
PROBE	0.987	0.987	0.984	0.99
R2L	0.981	0.966	0.827	0.828
NORMAL	0.789	0.704	0.948	0.949
Weighted Avg.	**0.943**	**0.786**	**0.98**	**0.981**

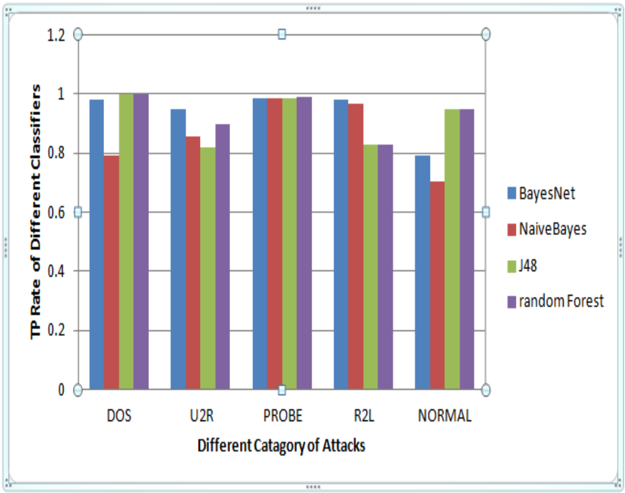


Fig. 3: Analysis of TP Rate with Different classifiers.

Table 2: Results of Precision of Each used WEKA Classifier

Class	BayesNet	NaiveBayes	J48	Random Forest
DOS	1	0.995	1	1
U2R	0.101	0.143	0.874	0.945
PROBE	0.562	0.088	0.992	0.997
R2L	0.568	0.432	0.812	0.813
NORMAL	0.987	0.97	0.952	0.953
Weighted Avg.	**0.968**	**0.948**	**0.981**	**0.981**

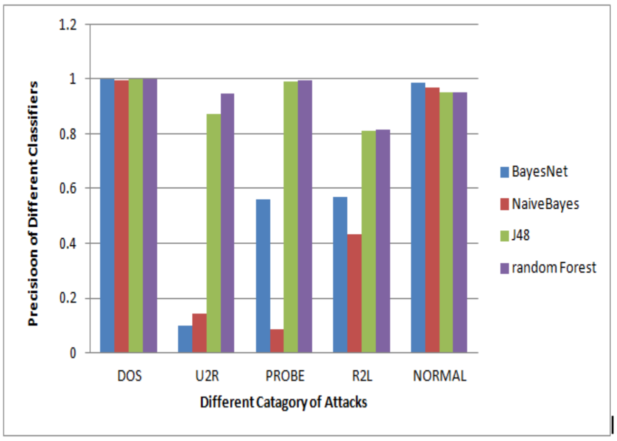


Fig. 4: Analysis of Precision with Different classifiers.

Table 3: Results of Recall of Each used WEKA Classifier

Class	BayesNet	NaiveBayes	J48	Random Forest
DOS	0.979	0.792	1	1
U2R	0.947	0.855	0.82	0.899
PROBE	0.987	0.987	0.984	0.99
R2L	0.981	0.966	0.827	0.828
NORMAL	0.789	0.704	0.948	0.949
Weighted Avg.	**0.943**	**0.786**	**0.98**	**0.981**

Table 4: Results of F-Measure of Each used WEKA Classifier.

Class	BayesNet	NaiveBayes	J48	Random Forest
DOS	0.99	0.882	1	1
U2R	0.183	0.246	0.846	0.921
PROBE	0.716	0.161	0.988	0.994
R2L	0.719	0.597	0.819	0.821
NORMAL	0.877	0.816	0.95	0.951
Weighted Avg.	**0.949**	**0.844**	**0.981**	**0.981**

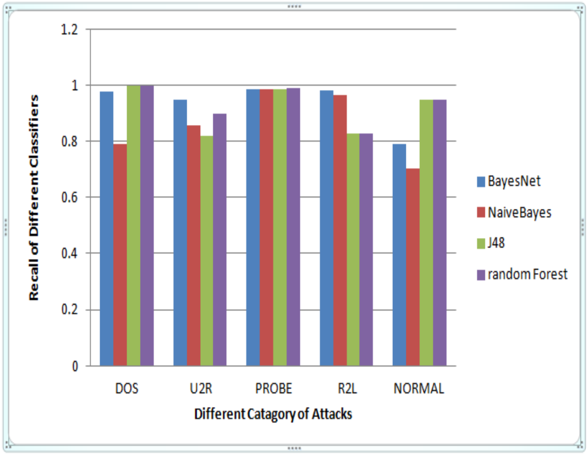
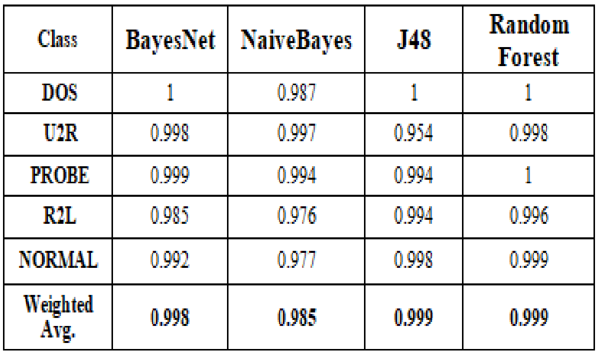


Fig. 5: Analysis of Recall with Different classifiers.

Table 5: Results of ROC of Each used WEKA Classifier



Class	BayesNet	NaiveBayes	J48	Random Forest
DOS	1	0.987	1	1
U2R	0.998	0.997	0.954	0.998
PROBE	0.999	0.994	0.994	1
R2L	0.985	0.976	0.994	0.996
NORMAL	0.992	0.977	0.998	0.999
Weighted Avg.	**0.998**	**0.985**	**0.999**	**0.999**

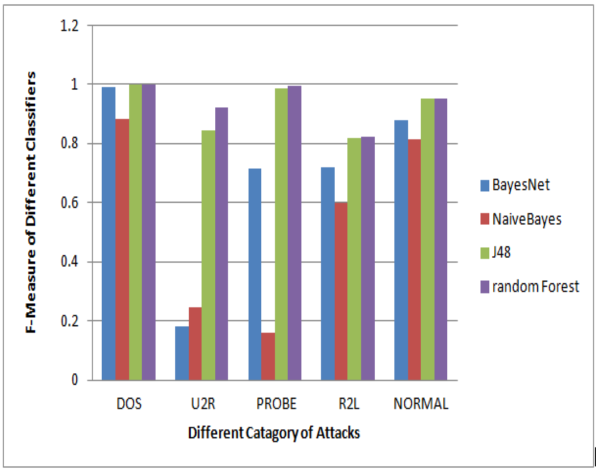


Fig. 6: Analysis of F-Measure with Different classifiers.

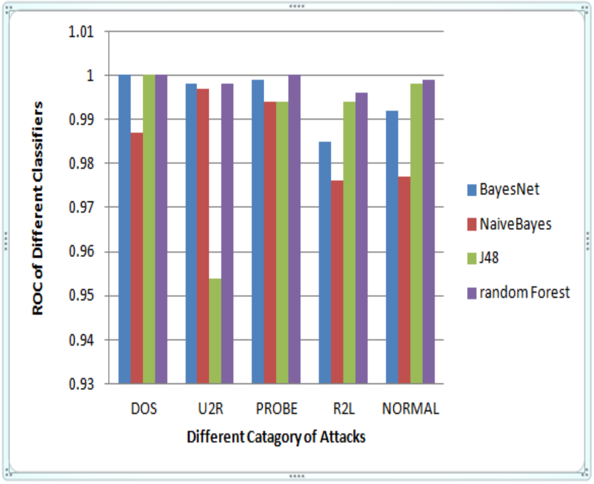


Fig. 7: Analysis of ROC with Different classifiers.

VI. Conclusion

An endeavor has been made to check the exhibition of the supervised machine learning Algorithms specifically BayesNet, NaiveBayes, J48, and Random Forest is looked at for intrusion detection. These calculations are evaluated with the KDDcup99 dataset. We proposed our ML-based strategy "A Machine Learning Based Intrusion Detection Framework using Kddcup 99 Dataset" with the assistance of the KDD CUP 99 dataset. The powerful classifier is recognized by looking at the exhibitions dependent on the Accuracy, Precision, recall, F- Measures, and RoC area. From the evaluated outcomes, it tends to be reasoned that the Random forest classifier beats different classifiers for the considered kddcup 99 dataset and boundaries. It has an accuracy of 98.1%. The work can be stretched out by considering the classifiers for multiclass characterization and thinking about just the significant credits for intrusion detection.

REFERENCES

[1.] Wang, Xiao-bin, Guang-yuan Yang, Yi-chao Li, and Dan Liu. "Review on the application of artificial intelligence in antivirus detection system i." In Cybernetics and Intelligent Systems, 2008 IEEE Conference on, pp. 506-509. IEEE, 2008.

[2.] Teresa F. Lunt., "A survey of Intrusion detection techniques", Computers and Security, Elsevier Advanced Technology Publications, 12(4):405-418, 1993.

[3.] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[4.] Emilie Lundin, Erland Jonsson" Survey of Intrusion Detection Research", Technical Report 02-04, Department of Computer Engineering, Chalmers University of Technology, 2002.

[5.] Lu, C-T., Arnold P. Boedihardjo, and Prajwal Manalwar. "Exploiting efficient data mining techniques to enhance Intrusion detection systems." Information Reuse and Integration, Conf, 2005. IRI-2005 IEEE International Conference on., pp. 512-517. IEEE, 2005.

[6.] Min, L. I. "Application of Data Mining Techniques in Intrusion Detection." A Yang Institute of Technology (2005).

[7.] Belavagi, Manjula C., and Balachandra Muniyal. "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection." Procedia Computer Science 89 (2016): 117-123.

[8.] Jianliang, Meng, Shang Haikun, and Bian Ling. "The application on Intrusion detection based on k-means cluster algorithm." In Information Technology and Applications, 2009. IFITA'09. International Forum on, vol. 1, pp. 150-152. IEEE, 2009.

[9.] Luo, Bin, and Jingbo Xia. "A novel Intrusion detection system based on feature generation with visualization strategy." Expert Systems with Applications 41, no. 9 (2014): 4139-4147.

[10.] Cao, Ning, and Yingying Wang. "A Novel Approach to Improve Robustness of Data Mining Models Used in Cyber Security Applications." In Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on, vol. 2, pp. 297-300. IEEE, 2017.

[11.] Bamakan, Seyed Mojtaba Hosseini, Huadong Wang, Tian Yingjie, and Yong Shi. "An effective Intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization." Neuro computing 199 (2016): 90-102.

[12.] Eesa, Adel Sabry, Zeynep Orman, and Adnan Mohsin Abdulazeez Brifcani. "A novel feature-selection approach based on the cuttlefish optimization algorithm for Intrusion detection systems." Expert Systems with Applications 42, no. 5 (2015): 2670-2679.

[13.] Aung, Yi Yi, and Myat Myat Min. "An analysis of random forest algorithm based network Intrusion detection system." In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017 18th IEEE/ACIS International Conference on, pp. 127-132. IEEE, 2017.

[14.] Elhag, Salma, Alberto Fernández, Abdullah Bawakid, Saleh Alshomrani, and Francisco Herrera. "On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems." Expert Systems with Applications 42, no. 1 (2015): 193-202.