*Turkish Journal of Computer and Mathematics Education Vol.12 No. 11 (2021), 2954- 2968*
*Research Article*

# A Detailed Analysis on NSL-KDD Dataset using various Machine Learning Techniques for Intrusion Detection

**Tarun Dhar Diwan, Dr. Siddartha Choubey, Dr. H.S. Hota**

[a]Chhattisgarh Swami Vivekanand, Technical, University, Bhilai, Chhattisgarh, India.
[b]Shri Shankaracharya Technical Campus, Bhilai, Chhattisgarh, India.
[c]Atal Bihari Vajpayee University, Bilaspur, Chhattisgarh, India

**Article History**: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

**Abstract:** Here we endorse for dynamic security defence one from the core technology is creation of Intrusion detection in the particular paper. Generation of a powerful device for security of network is possible by using this technology along with static security. At the first place, Devices with Intrusion technology enhancement is introduced in this paper. At the second place, networks and information generations speedy evolution, protection of network is at a critical stage and there are high complications in the issues of security of the network and the personal computers. Machine of Intrusion device new technologies are proposed in this paper. At last, Intrusion detection devices evolution is introduced. In a website the only need for any attacker is to insert a malicious code because of the net browser vulnerabilities in the field of security and protection aid and there may be a high chance of attack on the user that visits these wed sites. In the Intrusion detection research one of the main issues is to improve the overall performance by putting forward the devices to understand Intrusion detection system. To extend security property on the basis of prevention-primary introduction of Intrusion detection system is done. Violation of the devices security policy is defined as Intrusion..

**Keywords:** Internet; Intrusion; Security; Attack; Machine; Vulnerabilities; Threats

## 1. Introduction

With Intrusion issue little increment and net offering speedy evolution, with highly complicated intrusion nice painting with the technology of conventional intrusion detection is not possible[1]. Prevails number is getting higher in the net services, in the wed sites application are placed in large numbers, over the community through browsers might retrieve wed sites directly. As the private devices are generally not updated or patched, the factory of protection is the weakest in the user's personal computer[2]. It guarantees that the fine legal customers are allowed to get the right of entry to laptop belongings the application of machine analysing to pc safety, mainly to intrusion detection. It moreover describes the easy forms of intrusion detection mechanisms and their cutting-edge us of in attaining the goals of pc safety [3]. The task of detecting intrusions may be taken into consideration as a tool for getting to know missions as it involves the classification of conduct into patron and adversary conduct. System studying strategies towards fixing some hard computer protection problems in particular concerning detecting intrusions [4].

## 2. Tools in Intrusion Detection

More than some of the Organizations safety dreams an Instruction detection product contains present addresses. About security equipment’s are discussed in this phase.

### a. SNORT

Open supply and lightweight supply software program is a snort. To explain traffic the language that is based on run-primarily is used by the Chuckle. It saves the human readable shape packets for any IP address [5]. By content searching, protocol evaluation, vulnerabilities make the most attempt, detection of hundreds of worms is done by diverse per-processors laugh, different doubtable conduct and port scans [6].

### b. OSSEC-HIDS

Loose Open-Source type software program is OSSEC (Open-Source Security). It runs on the main running device and total architecture that is based on a server or consumer is used [7]. The capacity to serve OS logs to provide the service of garage and analysis is present in OSSEC. It is used and helpful in ISPs, engine, statistics facility, [8] efficient log evolution and university. Through HIDS analysis and monitoring of authentication log firewall is performed [9].

### c. FRAGROUTE

Terms like fragmenting router is fragroute's miles. IP packet is delivered to frag router from the attacker in this and then fragmentation is done and party is formed by conversion [10].

**d. HONEYD**

At the community digital hosts are created by this device called HONEYD. For network simulation on a LAN with more than one address by allowing a host request as the server are utilized by the host HONEYD [11]. To hint course, them or to knock the digital machine it has this capacity. Consistent with a document that has a configuration that is simple could be replicated by any kind of the resource on digital device [12].

**e. KISMET**

For WIDS (wireless Intrusion detection system) KIMET is a guiding principle. With packets happenings and payload of WIDS is consisted in WIDS [13]. It searches the bugler get admission to factory.

**3. Techniques Classification For Detection Specified Attack**

The categorization for distinctive attacks of various different techniques generally on the performance bases is provided by us in this last of the paper. For picking a method that is selected for the particular attack detection is facilitated to the readers by it. As a general dataset KDD' ninetynine is used for assessment by most of the methods [14].

**1. Detection of Denial of Service**

Attack Single-classifier method is much more simple to incorporate. With a detection rate of 97.24 percent, Decision Tree (DT) is doing well in this group. However, when 12 capabilities are used, the detection rate improves to 99.53 percent. With six characteristics, CANN achieves 99.99 percent accuracy. Whereas in Land assault detection it will perform admirably [15]. It wants to improve its set of features for other DoS attacks (in the observation [16] summary the explanation is explained). When ANN is paired with MARS and SVM [17], it increases efficiency and offers a classification fee of 99.96 percent. When combined with Fuzzy Clustering (FC-ANN), the efficiency of ANN is also enhanced, with a 99.93 percent detection rate. Fuzzy Logic along with ANN provides a detection rate of 99.5%. When we talk about SVM, with the Clustering method (CT SVM) when it is incorporated we can see that its detection rate increases [18]. Initially, just 91.6 percent was the detection charge, but it later increased to 97.35 percent. As a fact, when SVM is combined with ant colony networks (CSVAC) a added enhancement and up to 94.84 percent improvement is seen in the detection rate. All of the above-mentioned hybrid techniques take into account by 41 factors that influence time of computation with 23 features[19], SVM in combination with SA has the highest detection rate of a hundred percent, and with 19 capabilities it achieves 99.5 percent detection when combined with Hierarchical clustering. The use of the KDD'99 facts collection is one of the techniques. The categorization of many different techniques for detecting DoS attacks, along with their detection prices [20].

**2. Detection of Scanning (Probe)**

For Probe attack detection, Attack Single-classifiers with every 41 features do not show good performance. In this category, not even one of the classifiers achieves a score of 90%. The detection rate with 41 features of a Decision Tree is only 77.92 percent, but as it is combined with selected features (in number 12 features), it improves to 98.868 percent. Again a low detection rate of 36.65% is show by SVM but the detection rate improves to 86.46 percent as the consisting only eight features, MMIFS feature selection techniqueis uses. Increasing the integration of the single classifiers improves detection significantly. When paired with the Elman network, ANN achieves a detection rate of 100 percent where as the ANN alone provides a detection rate of 72.71 percent [21]. Even a detection rate of 99.79 percent is provided with the integration with MARS and SVM. Results on the bases of time of computing and detection rate is boosted by the Hybrid classifiers with feature selection. With 9 features, a combination of Subspace clustering, DBSCAN, and the EAR algorithm achieves a 100% detection rate [22]. With 12 features, the FNT technique with GA and PSO[23] attains a 97.89 percent detection score. SVM combined with DT and SA outperforms SVM alone, with a detection rate of 98.35% and 23 features, compared to 36.65% and 41 features for SVM alone [24].

**3. Detection of User to Root Attacks**

For U2R assaults, we have classified techniques solely on their detection fee. We've already listed the techniques which are working. The unmarried classifier's overall success could be catastrophic, in the situation of U2Rattack[23]. Whereas combining them with other classifiers improves their accuracy. With 12 functions, the FNT technique with PSO and GA [24] plays extensively and gets the 99.89 percent highest detection charge. For a particular form of U2R attack, wager password, the intrusion detection generation that uses Multiple Neural Network classifiers [25] has a 99.7% detection rate.The efficiency of Neural Networks has improved dramatically, with a detection rate of 94.01 percent when paired with the bushy clustering approach (FC-ANN) without an alternative within the collection of capabilities. ANN's detection fee for U2R attacks was previously zero percent. With increase in the detection rate to 67%, by the integration of ANN with SVM and MARS [26], but it is not right. When SVM (12%) is incorporate with the clustering method (CT SVM, 17.23%), there is very little growth,

which is also unacceptable. For U2R assaults, hybrid techniques with feature option have a high detection charge [27].

**4. Detection of Remote to User Attacks**

We divided techniques into categories based on the cost of detection for R2L attacks. The strategies that may be effective are discussed. The efficiency of the unmarried classifier may be very low in the case of R2L attacks [28]. The combination of different classifiers, on the other hand, increases their overall efficiency. As claimed by the author, the ensemble of SVM, ANN, and MARS achieves a 100% detection rate, which is a significant betterment over ANN (26.72 percent detection rate) [29]. They have, however, stopped mentioning which R2L assaults have been identified. A hybrid classifier with feature selection is also gaining popularity. With 12 features, the FNT technique with PSO and GA offers a detection rate of 98.99 percent. With 16 capabilities, FPSO achieves a detection rate of 98 percent. With 23 capabilities, the combination of SVM, DT, and SA achieves a detection rate of 90.78 percent. Other techniques' overall results. Rather then U2R attack better performance is shown by Multiple classifiers with feature option for R2L attack detection [30].

**This is attributed to the following:**

(1) Most of the hybrid classifiers are filtering the data before training the classifier by performing clustering on the data which group data items into groups based on their similarity measure. It brings the balance in data which is very crucial in machine learning algorithms [31].

(2) Multiple classifiers are more adaptable than single classifiers and hence can learn the new attack behaviour efficiently. However, still, there are various difficulties associated with detecting low-frequency attacks as

Current intrusion detection approaches focused on machine learning have been extensively investigated using recognize-to-person attack categories. With the potential solution, with limitations of the processes for each type are discussed, along with potential solutions [32]. There is no single specific gadget learning algorithm that can assist in identifying all types of problems. As a result, the application of a specific set of rules (anomaly, misuse, or h) is required. [33].

**4. Issues In The Attack of Low-Frequency Detection**

The attack record gathered from the data statistics of Machine Learning Algorithm works. With the keen inspection of the data set extracted of the inclined host machine connection detection of the Prode and DoS attacks might be done without difficulty, whereas the low frequency attacks that consists even R2L and U2R by carefully examining the dataset of the connection with the help of KDD'99 datasets detection is difficult. The following are the reasons for so:

**(i)** Low-frequency attacks have linked statistics that are very close to the regular connections.

**(ii)** The behaviour of R2L and U2R link records is very similar. As a result, it's difficult to tell the difference between U2R and R2L attacks. As a fact, one of the variant of the R2L attack is the U2R attack. A consumer does not have local access to the computer in an R2L attack. He must first gain access to a regular user's account using various account hijacking exploits in order to gain root privileges [34]. The attacker may then initiate more exploits to obtain root privileges after logging in as a regular user, while in a U2R attack, the attacker has unprivileged local access to the victim [35].

**(iii)** A single link may be used to initiate the attacks of low-frequency. The relation information given by the KDD'99 dataset is insufficient. While some of the Content attributes, such as the root shell (P14), root shell (14), num compromised (P13), number of failed logins (P11) and so on, are present in the KDD'99 dataset, they are insufficient for attack detection [36]. The load module attack (U2R), for example, In the directory /dev generates special devices and from the current system two kernel drivers that are loadable dynamically are loaded to access the modules. An unauthorised user can gain root access on the local machine due to a bug in the way the load module sanitises the environment. The attack can be identified by looking for 'load module' and the strings set $IFS='V in the user's session [37]. Using machine learning algorithms and the KDD'99 dataset, this form of keyword spotting is difficult to attain.

**(iv)** When compared to Probe and DoS attacks, the number of R2L and U2R samples in the KDD'99 testing and training dataset is very small. The classifier is less appropriate to detect such attacks due to inadequate learning of such attacks [38]. Furthermore, the classifier treats such attacks as natural due to the uneven distribution of data.

**(v)** The number of amount of operations performed as root, root shell login, files created, and other activities performed by these attacks may be similar. Detecting attacks of low-frequency becomes more difficult in this situation. However, a thorough analysis of the system call traces for invocation of specific commands [39], suspicious sequences of system calls, the existence of specific modules or processes and other indicators of attack

operation in the system could provide some clues. The FFB config attack, for example, takes advantage of a buffer overflow (U2R). It sets up the Creator Fast Frame Buffer (FFB) Graphics Accelerator, that is section of the SUNWffbcf software package for FFB Configuration. The identification of the attack can be done by finding the command '/usr/sbin/ffbconfig/' with an overweight argument for the '-dev' parameter in the device call traces [40].

**(vi)** While some approaches to detecting certain attacks achieve a high level of accuracy (around 90.99 percent), we cannot guarantee that these techniques will offer the same level of precision for identifying unknown threats or freshly created R2L or U2R attacks. Because the methods have been verified against the KDD'99 test database, that includes the feature values of the attacks, they may be enough to distinguish them from Probe and DoS. A dictionary attack, for example, is an R2L invasion in this the attacker guesses usernames and passwords repeatedly in order to obtain remote access to a computer [41]. The attack can be identified by looking at 2 factors: the number of failed login attempts (P11) over time and each service's session protocol (P2). However, if the feature values do not provide sufficient details, as is the case if the victim's password is weak and the attacker gains access to the victim's computer in one or two attempts, like by using [42] his school name or phone no and similar. This attack's feature values will be identical to those of a regular link. The machine learning algorithm will not be able to identify these attacks in this situation. Low frequency attacks are difficult to detect simply by looking at network features [43]. The challenges of detecting low-frequency attacks were identified and addressed. The potential viability of solutions to identify attacks including virus, dictionary attack, password cracking, buffer overflow, etc. has been addressed. The latest works on identifying these attacks using analysis by device call [5] can be found here. In one of the recent paper, we looked at both network features and system call to detect low-frequency attacks.

**5. Challenges and Needs and Objective**

There is various demanding system to enforce an IPS device. To unmatched point of failure and painting inline offline offering a choke point with a high potential this IPS tool is designed. If whilst the inline gadget fails by the impact of the community performance and passive IDS fails in this situation some of the attacks go undetected that are purchased. Some of the community component tend to carry out NIPS (Network Intrusion prevention machine) such as community switch [6]. To organize the demanding situation with the help of it by meeting the overall reliability and performance needs of the community, that results only some customer inclination to sacrifice community reliability and performance for the purpose of protection [7].

Visitors are [42] slowed down by the NIPS and the NIPS problem of dropping packet. Information circulating can be performed with the help of NIPS. With superior ASICs and FPGAs using custom hardware the trouble of maximum excessive give up is shown in IPS vendors. For functioning as Intrusion prevention and detection device it is very important to layout product. IDS that is a kind of defence system that is the desire of all employees [22]. Cooperation tackles whilst at the duration of development of an Intrusion device but there are some challenges in this.

**(A)** Few upgrades are still going on in the IDS era. It is very essential for any employee it was understood by the IDS implementation. No human intentions are needed in IDS technology. Some of the automations such as for a specific given time the malicious connections are shunned, in the situation any malicious activity is found notifying the administrator and to control listing which will forestall a connect that might be [8] malicious by dynamic enhancing router's get entry [13]. Mentioning the IDS log is mandatory for occurrence of all events. The functions that are detected with the help of IDS over some time is analysed with the help of tracking logs based on everyday foundation [16].

**(B)** On the fulfilment of the development implementation of IDS are based. For implementing and designing phase it is very crucial to make plans. It is perfect to put in force host-based totally IDS and for a community primarily based hybrid solution. Amongst agencies the variations of decisions are possible. For most of the agencies instantaneous choice is a network-based IDS and this is due to the capability [17] of screen multiple system and additional truth doesn't needs a software program that has to be loaded on manufacturing machine in contract to host-primarily basis IDS. Hybrid answers are provided by few of the groups. Hence before putting in bunch-based sensor for a system available resources are wished [18].

**(C)** There is a need to acclimate the sensor supervisor ratio. Before beginning the IDS implementation, it is very necessary for the baseline policy to be designed beforehand and avoiding false positive end result [36]. Shipping some of the positive end result to the sensor can also be shipped by the IDS sensor and insufficiency of ratios is also possible [27].

**(D)** In place of being proactive IDS is still very reactive and on the assault signatures this technology works. Define in advance samples of attack are described in signatures [18]. An exclusive form of attack is found and they are constant in frequency and database of signature replace differs from supplier to supplier hence the signature database wishes to be up to date on every occasion. From the other port traffic in and out port can't be

seen in the switched network due to the collision domain names. It might be regarded for another port for network inside and outside in HUB based totally network. There is a requirement to come across site visitors inside and outside of the port by the NIDS sensor and inside the environment of the switch for the malicious visitor [24]. Port spanning or mirroring is used for reaching this type successfully.

**6. Preparation of Data**

Protocol sort, standing flag and provider type are basic capabilities of KDD, while the remaining 38 functions are non-stop functions. The frequency of different groups that existed for certain functions was used to translate categorical features into non-stop functions. Since the functions in KDD are of different natures and their scales are of excessive variance, normalisation is a required step in the training of the education datasets. The proposed equations 2 and 3 were used to normalise the dataset [2].

The normalized value of feature i, NVi is computed as follows:

$$NVi = \text{normalize}(\ln(\text{vali} + 1)), \quad (1)$$

$$\text{Normalize}(xi) = xi - \ln(M\,ini + 1)\ln(M\,axi + 1) - \ln(M\,ini + 1), \quad (2)$$

Here vali is the function I preliminary fee, and Mini and Maxi are the function I minima and maximum values, respectively. To test the tolerance of supervised techniques on the distribution of education statistics, datasets with exceptionally large relative populations of everyday and assault data were generated using the population categories mentioned. In each population group, ten unique TS were chosen [2,3]. The facts were chosen at random from the normalised dataset, but the selection was made in such a way that each TS included all four attack groups [18].Four attack groups were consecutively organized. To keep away from this hassle, in every TS preparation, a extraordinary percentage of each assault class turned into taken into consideration [1].

**A. Experimental Results: 1**

The strategies had been carried out to every T Si. Similar test facts become used for special T Si in every category. The outcomes of every approach for 10 one-of-a-kind TS (in each class) had been combined to evaluate the approach. In the primary evaluation, we have taken into consideration the most detection charge for exceptional dataset categories from distinct strategies [10].

In Table 1 the consequences are represented. For the unique population class price of most detection is almost same for ever technology. In special strategies. In the manner in which distribution of attack inside dataset no longer affects max detection fee at special strategies.

**Table 1** Maximum Detection Rate

| **Attack Category** | | **DoS** | **Probe** | **R2L** | **U2R** |
|---|---|---|---|---|---|
| Gaussian | 8020 | 0.967 | 0.873 | 0.136 | 0.486 |
| | 8416 | 0.967 | 0.874 | 0.136 | 0.486 |
| | 8812 | 0.843 | 0.86 | 0.135 | 0.486 |
| Naïve Bayes | 8020 | 0.791 | 0.816 | 0.125 | 0.843 |
| | 8416 | 0.791 | 0.814 | 0.125 | 0.814 |
| | 8812 | 0.791 | 0.813 | 0.124 | 0.843 |
| Decision Tree | 8020 | 0.972 | 0.736 | 0.093 | 0.286 |
| | 8416 | 0.968 | 0.779 | 0.074 | 0.243 |
| | 8812 | 0.966 | 0.747 | 0.09 | 0.3 |
| Random Forest | 8020 | 0.972 | 0.745 | 0.055 | 0.257 |
| | 8416 | 0.971 | 0.758 | 0.048 | 0.343 |
| | 8812 | 0.971 | 0.777 | 0.06 | 0.257 |

Table 1. summarises the appropriate approach for various attack groups based on the maximum detection charge. The results of the Decision Tree and Random Forests indicate that DoS has a detection rate of over 96 percent. For U2R, Naive Bayes has a detection price of over 84 percent. Any strategy has a negative impact on the R2L group. Within the Gaussian technique, the nice end result is 0.136 [2]. For the Probe assault with the detection rate greater then 81% for Gaussian and Naive Bayes. The same deviation is calculated inside the price of identification for T Si distinctively in all classification. In calculations [1].

**Table 2** Mean and Standard Deviation Rate

| **Attack Category** | | **DoS** | | **Probe** | | **R2L** | | **U2R** | |
|---|---|---|---|---|---|---|---|---|---|
| **Technique** | Population Category | Mean | Std | Mean | Std | Mean | Std | Mean | Std |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Gaussian** | 8020<br>8416<br>8812 | 0.967<br>0.967<br>0.828 | 0.0<br>0.0<br>0.0 | 0.866<br>0.848<br>0.844 | 0.002<br>0.012<br>0.004 | 0.135<br>0.135<br>0.135 | 0.0<br>0.0<br>0.0 | 0.461<br>0.461<br>0.457 | 0.006<br>0.017<br>0.011 |
| **Naive Bayes** | 8020<br>8416<br>8812 | 0.791<br>0.791<br>0.791 | 0.0<br>0.0<br>0.0 | 0.805<br>0.803<br>0.866 | 0.005<br>0.005<br>0.004 | 0.097<br>0.097<br>0.097 | 0.03<br>0.03<br>0.03 | 0.766<br>0.759<br>0.761 | 0.043<br>0.039<br>0.035 |
| **Decision Tree** | 8020<br>8416<br>8812 | 0.887<br>0.926<br>0.509 | 0.066<br>0.058<br>0.303 | 0.681<br>0.708<br>0.678 | 0.024<br>0.025<br>0.036 | 0.054<br>0.054<br>0.05 | 0.016<br>0.014<br>0.024 | 0.171<br>0.161<br>0.17 | 0.066<br>0.034<br>0.037 |
| **Random Forest** | 8020<br>8416<br>8812 | 0.827<br>0.598<br>0.512 | 0.186<br>0.31<br>0.266 | 0.718<br>0.699<br>0.712 | 0.011<br>0.025<br>0.037 | 0.033<br>0.023<br>0.037 | 0.007<br>0.011<br>0.008 | 0.183<br>0.199<br>0.174 | 0.052<br>0.063<br>0.058 |

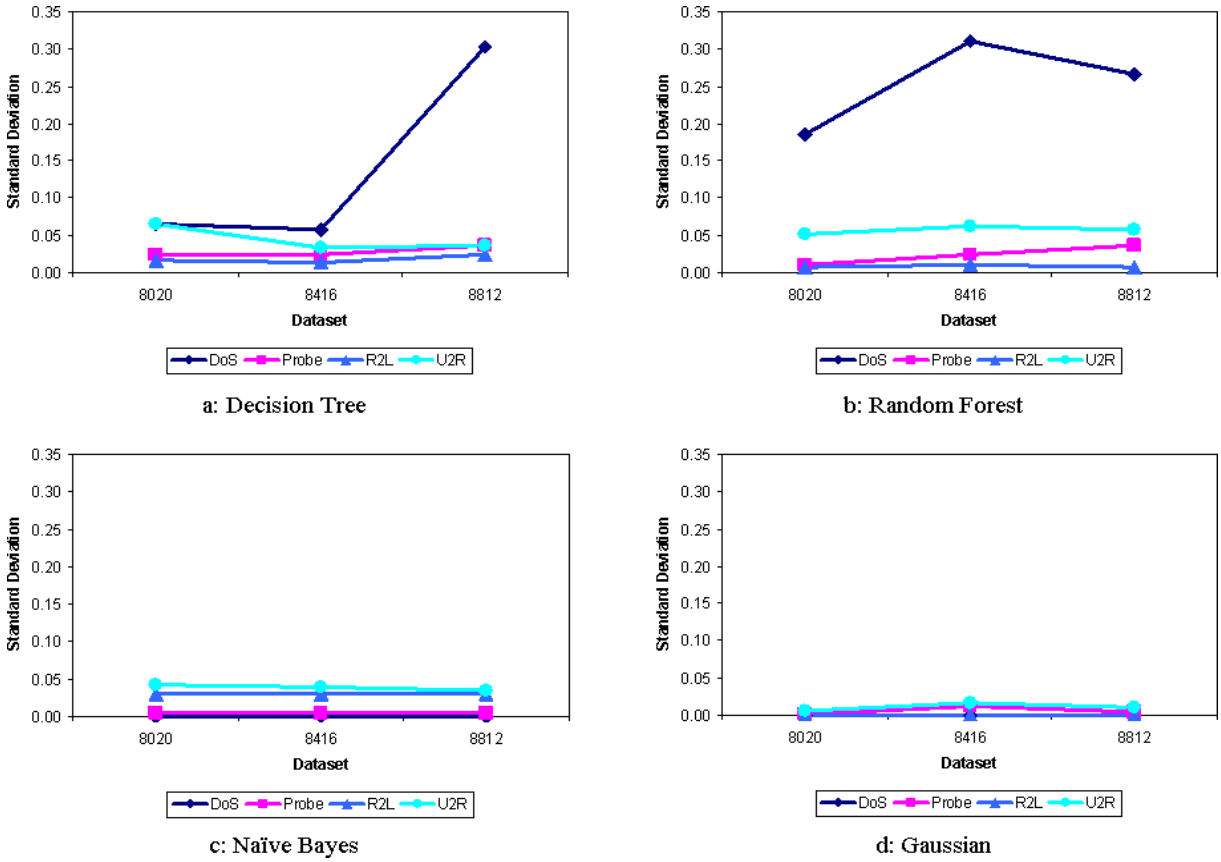


**Figure 1.** Standard Deviation in Detection Rate

Within the system, we ignored the most plus minimal detection costs of T Si within each populace course to maintain away through the impact associated with outliers. The recommend and the regular even method's deviation in distinct classes of the data set [3].

Figure1. Shows that the Choice Tree has a high standard, particularly when it comes to DoS recognition. Despite the fact that Decision Woods has an exceptionally high maximum rate for DoS detection, it is extremely sensitive to the two populations and sections. The normal results for the Arbitrary Forest technique will be shown in Figure. The sensitivity of this approach to people who have been attacked would be worse than Option Tree. Figure shows the Naive Bayes method's conclusion result, which shows low standard deviation in one-of-a-kind attack courses. Figure displays the effects associated with the Gaussian method [1]. The particular approach has a particular first-class deviation outcome between all the methods.

**(B) Experiment – 2**

EVALUATION METRICS

Each of these assessment parameters are derived from the confusion matrix's four basic attributes, which represent the real and expected groups [1].

1. 1) TP / True Positive - As attack the rightful categorization of the attack data.
2. 2) FP / False Positive - As attack the wrongful categorization of the general data.
3. 3) TN / True Negative - As general the rightful categorization of the general data
4. 4)FN / False Negative - As general the wrongful categorization of the attack data

To assess the success of our proposed solution, we will use the following measurements:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The accuracy is the percentage of correct classifications out of a total number of correct classifications.

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

The accuracy is calculated by dividing the number of correct classifications by the numberof incorrect categorization.

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

The recall (also known as detection rate) is a metric that measures the number of correct classifications divided by the number of incorrect classifications.

The F1score is a derived effectiveness measurement that calculates the harmonic mean of precision and recall.

**7. Test Evaluation.**

These particular phase notes the particular experimentation procedure plus outcomes received. All of the tasks are completed using Python along with sci-package-analyze and tensor-go with the movement library at the particular Windows 10 system. The test laptop is equipped with an Intel(R) Core i7 CPU running at 1. 8GHz and 8GB of RAM. 1) EVALUATION OF THE TEST To begin, a variety of pre-determined system learning algorithms are used to move-validate and analyse the indicators of each set of rules [1]. The genuine training set is divided into two parts: education and verification, in a 50:50 ratio.

**Table 3** on the training set Cross-validation.

| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **FI** | **Time(s)** |
|---|---|---|---|---|---|
| DeciTree | 99.63% | 99.62% | 99.23% | 99.25% | 0.33 |
| RainForest | 99.8% | 99.7% | 99.8% | 99.79% | 0.7 |
| KNN | 99.59% | 99.57% | 99.59% | 99.50% | 33 |
| LR | 97.73% | 97.72% | 97.73% | 96.66% | 13.7 |
| SVM | 99.53% | 99.52% | 99.53% | 99.48% | 220.7 |
| DNN | 99.09% | 99.09% | 98.4% | 98.64% | 245.2 |
| Adaboost | 99.9% | 99.9% | 99.9% | 99.45% | 112.3 |

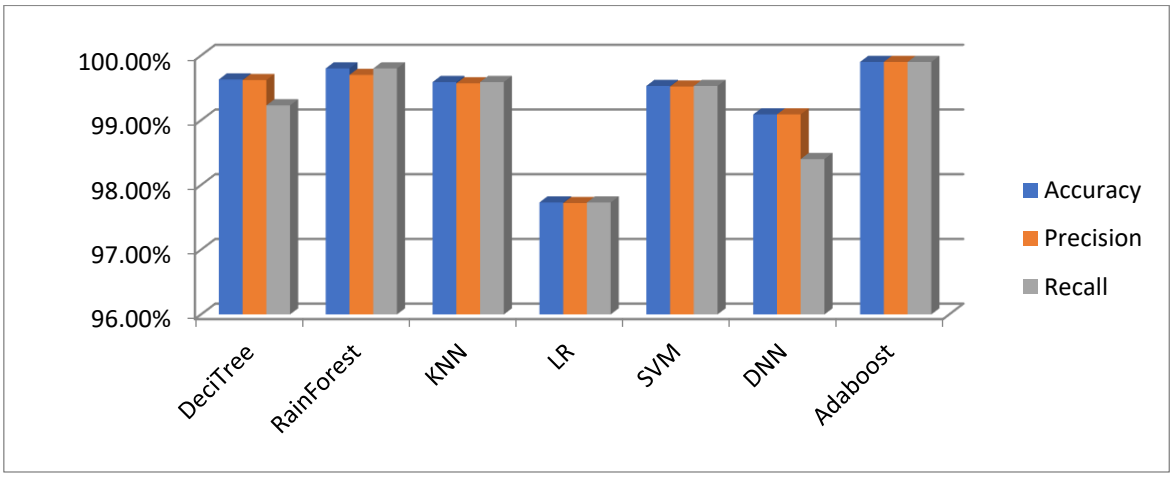


**Figure 2.** graphical representations of Cross-validate on training set

The education impact of the Adaboost set of rules is good, according to the validation results in Table 3. For the validation of the quantification and algorithms of the effect of generalization in the 4th table the usage of the test data set is done.

**Table 4** Result of each algorithm on KDDTest+

| Algorithm | Accuracy | Precision | Recall | FI | Time(s) |
|---|---|---|---|---|---|
| DeciTree | 79.63% | 83.62% | 79.23% | 77.25% | 6.23 |
| RainForest | 77.8% | 79.7% | 76.8% | 73.79% | 1.86 |
| KNN | 76.59% | 80.57% | 76.59% | 71.50% | 33 |
| LR | 74.73% | 80.72% | 97.73% | 70.66% | 13.7 |
| SVM | 79.53% | 82.52% | 79.53% | 69.48% | 220.7 |
| DNN | 79.09% | 89.09% | 78.4% | 82.64% | 245.2 |
| Adaboost | 76.9% | 83.9% | 78.9% | 73.45% | 112.3 |

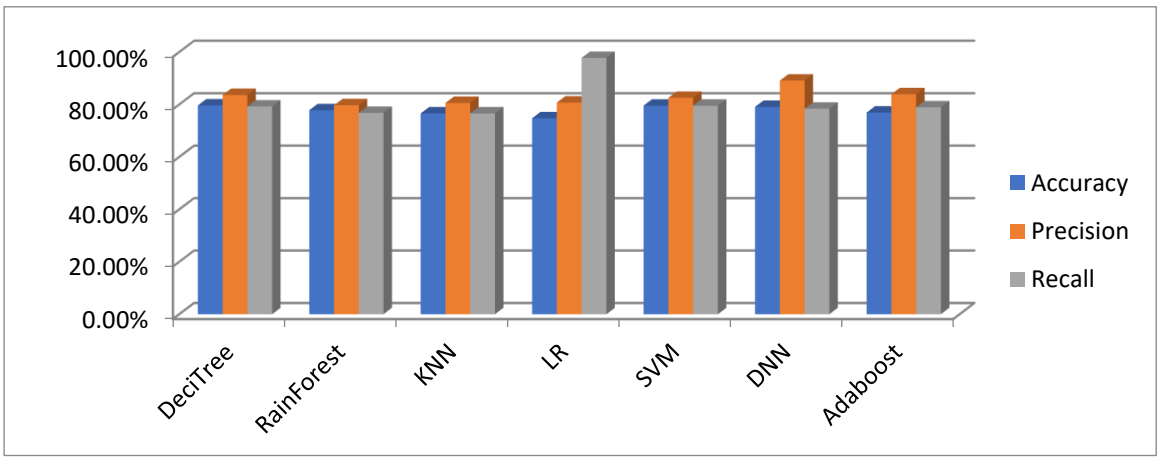


**Figure 4.** Result of each algorithm on KDDTest+.

In intrusion detection field, an enormous amount of real-time is needed for the analysis of large volumes of security log data, which necessitates not only high precision but also as fast a detection time as possible. In comparison, it has been determined that the overall accuracy is higher of the decision tree and DNN, and the selection tree's walking time will be shorter, that comes under the fine cost-effective studying set of rules [1, 2]. We also need to evaluate the detection effect of every set of rules for one-of-a-kind types of information, the detection effect of R2Land U2R information will be terrible, that relates very nearly to the pattern imbalance share within earlier examination. If all of us want to enhance the general detection effect, we need to locate ways to triumph over these issues. Even though DNN has a better overall detection effect, it performs significantly worse

in Normal-type detection; some different algorithms work well on Normal-type, but not on the others [1, 12]. As a result, various classification algorithms do not gain in most situations, but each has its own set of advantages. In the future, we will definitely refine the algorithm mix and combine the advantages of multiple algorithms to increase the overall detection effect [11].

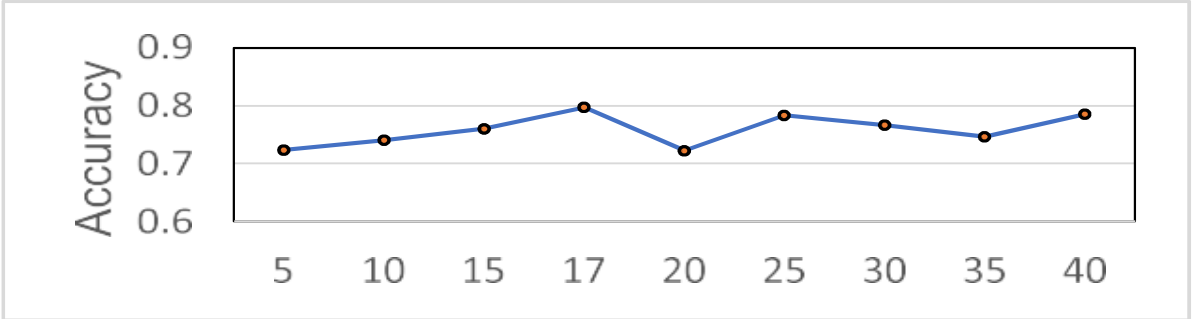


**Figure 5.** By the usage of the various number of features for decision tree accuracy.

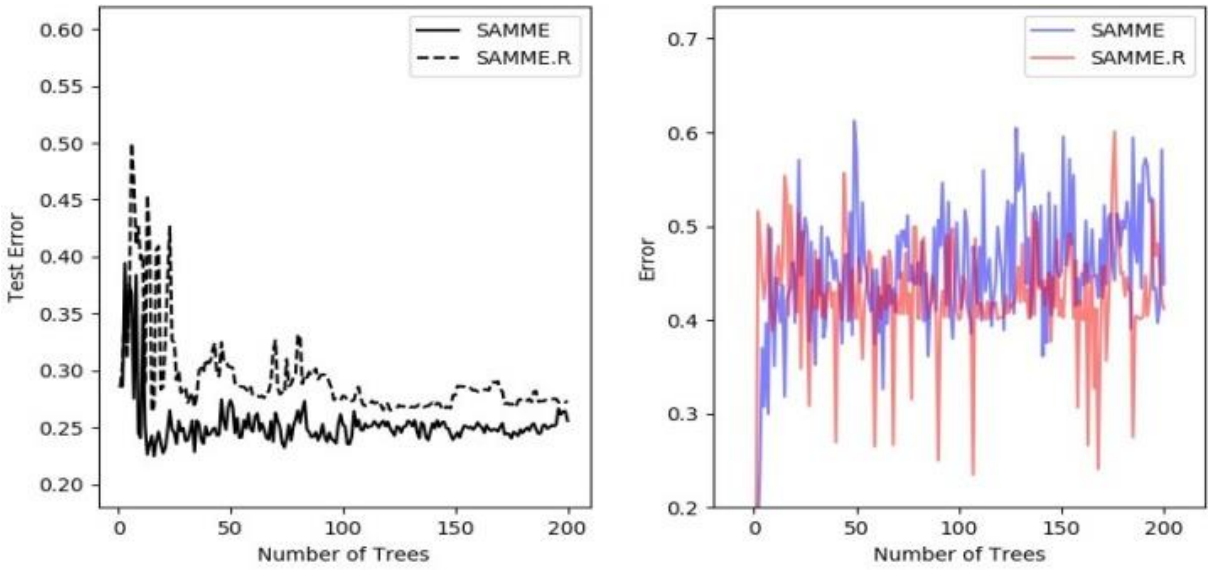


**Figure 6.** Adaboost SAMME.R test results

**Multi Tree Result**

The role selection approach is used by many researchers to boost the decision tree's effectiveness. The partitioning functionality is chosen based on the Gini function with the highest cost. In Figure 8, we use the CART (Classification and Regression Tree) set of rules to test the accuracy of various numbers of capabilities, and we find that 17 functions have an NSL-KDD Test+ accuracy of 80. 21%. We looked at the four-stage decision tree and found that the src bytes function has the highest Gini value and is the first-class choice for the root node in Figure 6. Experiment show that the number of functions and features chosen have a large influence on the detection performance. We may also choose seventeen primary features for the education of the specific decision tree in the following set of rules [1].

The Adaboost SAMME. R algorithm teaches many vulnerable classifiers to form a powerful classifier [14], but the recognition impact isn't as good as the precision of using choice tree group of guidelines, and the precision of using two hundred estimators isn't as good as the precision of using choice tree group of guidelines, as shown in Figure 9, suggesting that the Adaboost algorithm isn't generally effective [1]. All of us make use of the NSKDD Check information set in order to compare the choice woods, Adaboost, and Multitree set of guidelines. The multiTree all of us proposed achieves the particular first-rate impact, as well as accuracy is 84. 23% [3].

**Adaptive Voting Result**

Following the test and analysis above, the SVM formula takes a long time and has little advantage in terms of accuracy, as the Adaboost algorithm isn't flawless, and logistic regression's particular precision [1]

**Table 5** comparison in between Multi Tree and Adaboost.

| **Algorithms** | **Accuracy** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|
| Decision Tree | 79.71% | 83.51% | 79.72% | 77.31% |
| Adaboost | 76.02% | 81.82% | 76.02% | 72.12% |
| Multitree | 84.23% | 86.4% | 84.23% | 83.6% |

Give up these three algorithms because the specific algorithm isn't very good. Decision Tree, Random Forest, k-NN, DNN, and MultiTree are chosen as ensemble learning algorithms based on detection accuracy and overall operation efficiency. Table 9 shows a few examples of how the adaptive balloting set associated with rules functions, as well as what the numbers on the desk mean in terms of sample types. The weighted balloting group of rules can be used to determine the specific final prediction results after summing up the results associated with each formula [3]. It's worth noting that this voting formula has a higher degree of accuracy than single-arranged rules. The adaptive vote casting arrangement of rules can be used to validate the particular NSL-KDD look at the particular set after training with various algorithms, and basic outcomes are determined from Desk 6 [1].

**Table 6** Result of voting

| **Method** | **Normal** | **Dos** | **Probe** | **R2L** | **U2R** |
|---|---|---|---|---|---|
| Voting | 94.93 | 84.37% | 87.11% | 55.27 | 25% |

**Table 7** Adaptive voting sample on KDDTest+

| DeciTree | R.forest | kNN | DNN | Multi Tree | **Voting** | **True** |
|---|---|---|---|---|---|---|
| 3 | 0 | 0 | 3 | 0 | **0** | **0** |
| 1 | 1 | 1 | 1 | 1 | **1** | **1** |
| 4 | 0 | 4 | 4 | 3 | **4** | **4** |
| 3 | 4 | 0 | 4 | 4 | **4** | **4** |
| 4 | 0 | 4 | 4 | 2 | **4** | **4** |
| 0 | 0 | 0 | 0 | 2 | **2** | **2** |
| 3 | 0 | 0 | 4 | 2 | **4** | **3** |
| 0 | 0 | 0 | 4 | 3 | **3** | **3** |
| 2 | 2 | 2 | 3 | 2 | **2** | **2** |
| 0 | 0 | 0 | 0 | 3 | **3** | **3** |
| 0 | 0 | 0 | 4 | 3 | **3** | **3** |
| 2 | 0 | 2 | 2 | 0 | **0** | **0** |

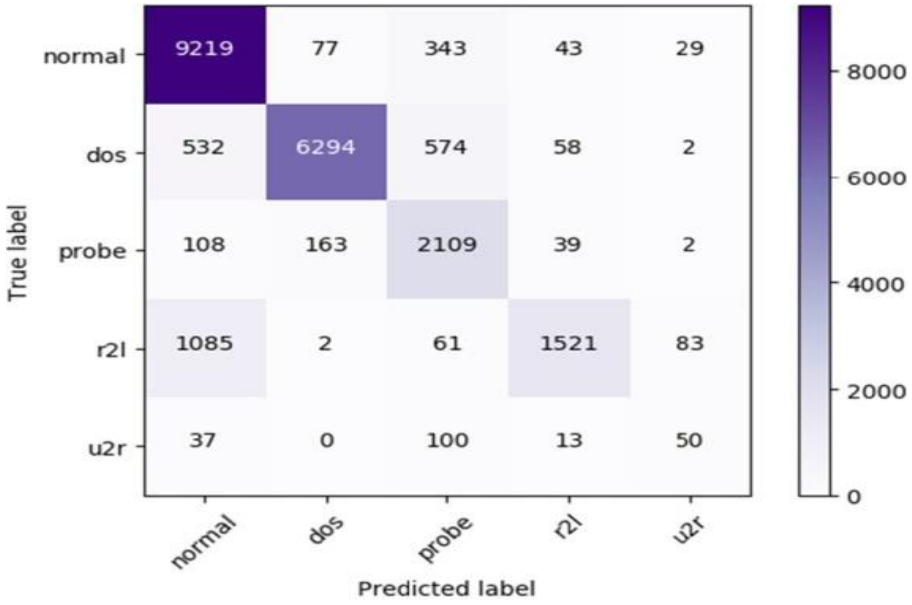


**Figure7.** Confusion matrixes of adaptive voting (85.2%)

Figure7.The Accuracy of metrics: 0.852, Recall: 0.852, Precision: 0.865, F1: zero.849.

The PCA major issue evaluation technique is used to investigate the take a look at end result [3] proven in Figure 8. It has been discovered that the documents distributed within the examination collection have certain parallels to educational statistics. R2L and U2R data, for example, overlap with various statistics and have a non-uniform distribution, making classification difficult. Next to the use of the enhanced set of rules to categories take a look at records, the misguided categorized records are displayed through the PCA technique. Most of the facts points inside the test records had been efficiently labeled and deleted from the vote casting results determine [2]. Normal and Probe data are clearly divided, but a few DOS (red) statistics are incorrectly defined. However, it also implies a linear normal distribution, which helps one to keep optimising [1].

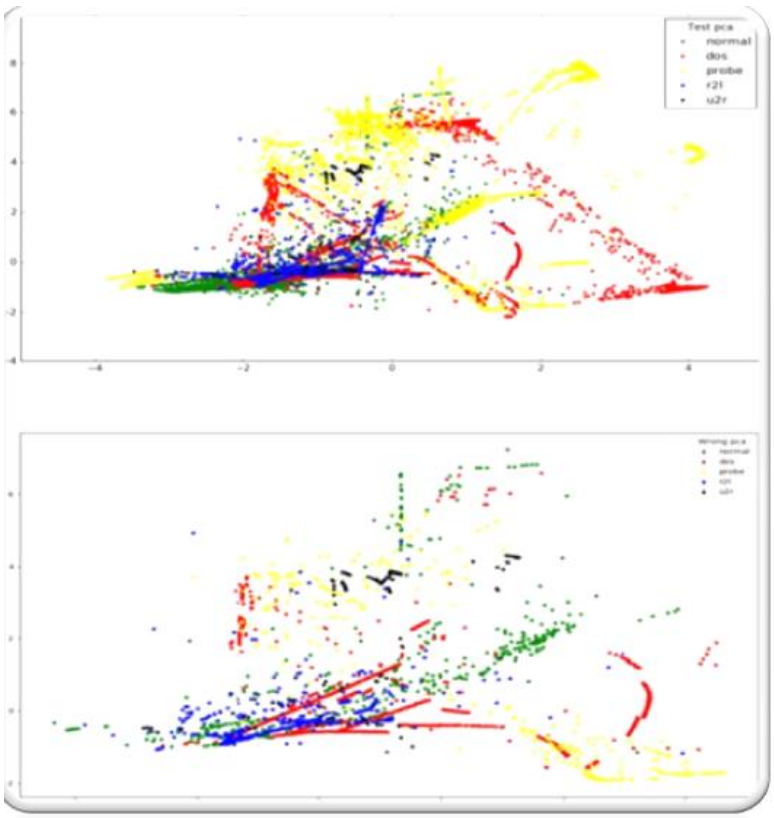

**Figure 8.** With the voting wrong result PCA compare test set (14.8%).

**Table 8.** Along with the other model comparison of our model.

| Author | Algorithm | Classes | Data Set | Accuracy |
|---|---|---|---|---|
| Our Model | Ensemble Voting | 5 | KDDTest+ | 85.2% |
| Our Model | Multi Tree | 5 | KDDTest+ | 84.23% |
| Our Model | DNN | 5 | KDDTest+ | 81.61% |
| Majdlatah | KNN+ ELM | 5 | KDDTest+ | 84.29% |
| KEHEWU | CNN | 5 | KDDTest+ | 79.48% |
| Tavallaee | NB TREE | 5 | KDDTest-21 | 66.16% |
| Ingre B. | ANN | 2 | KDDTest+ | 81.2% |
| Aggrawal P | Random Tree | 5 | KDDTest+ | 83.04% |
| Ambusaidi | LSSVM-IDS | 5 | Corrected KDD 99 | 78.86% |
| Al-Qatf M | SAE_SVM | 2 | KDDTest+ | 84.96% |

**Comparison of performance**

We review the test results with statistics from different articles [15] in order to critically assess the impact of our collection of policies. Table 8 shows that our adaptive ensemble method examining the application is a successful strategy for intrusion detection, and our ensemble version provides a high-quality attack class on the KDD Test+ dataset [1,3].

**8. Concusion and Future Work**

In this paper, supervised intrusion detection techniques are evaluated using training datasets with various attack populations and percentages. The simulation results demonstrate that in each attack group, the overall detection rates for the three population groups are the same. Random Forests and Decision Trees perform well in detecting DoS in the maximum detection rate study, while Naive Bayes and Gaussian perform better in these other assault classification [2]. For each population group, we also looked at the standard deviation of the detection rate for various techniques [3]. On the basis of databases with varying assault percents, the methods' efficiency is assessed. The Gaussian technique has the least sensitivity to various training datasets in this experiment. Naive Bayes has a greater sensitivity for U2R and Probe attacks than Gaussian [1]. Random Forests and Decision Trees are also shown to be highly sensitive to various attack categories, especially DoS attacks [3]. When tested utilizing various training datasets, probabilistic techniques demonstrate more robustness than predictive techniques, according to the findings of this paper. It's also been noted that predictive approaches are successful. Probe, U2R and R2L have a higher detection rate in data with fewer samples. [8].

**Table 9** Proper Techniques for Different Attack Categories

| Technique | Attack |
|---|---|
| Ttaussian | R2L, Probe |
| Naïve Bayes | U2R, Probe |
| Decision Tree | DoS |
| Random Forest | DoS |

Apart from potential portraits, we want to find the best mix of these intrusion detection techniques. According to the results, Random Forests and Decision Trees are good candidates for spotting DoS, whereas Naive Bayes and Gaussian are great for detecting various forms of attack [4]. In addition, while Naive Bayes and Gaussian are less sensitive to training datasets, Random Forest and Decision Tree are more sensitive [2]. We assume that integrating these strategies properly would result in a less sensitivity and higher detection rate to the statistical distribution [3]. Customer safety and privacy have been harmed as the number of intrusions into the network and host machines has increased. Numerous solutions to detect intrusions have been developed by researchers [2]. In our paper, we looked at the security aspects of intrusion detection using a device studying approach. With a brief overview of their attack capabilities, we have identified various styles of attacks within the group and host structures [2,3]. The research found that if a system works well for detecting one attack, it will not work as well for detecting other attacks. The critical overall output evaluation of a variety of gadgets learning algorithms was

carried out in an evolutionary manner. The distinction was made between single classifier processes and multiple classifier methods [14]. The influence of a classifier on another classifier is not only investigated, but also the influence of a function subset on the classifier. We've shown that while an ultimate feature set is sufficient for analyzing an assault's actions, it is insufficient for analyzing the behavior of various attacks. As a result, there is a need to define the gold standard feature subset as well as a suitable approach for and type of assay. The difficulties involved with low-frequency attacks indentification using device learning strategies across network data sets were identified. It allows researchers to work on other options in the hopes of coming upon Low-frequency attacks [19]. Recommendations for future research are given to assist researchers in their quest for more efficient assault identification solutions. To generalise our findings, we described current literature that is primarily based on comparable techniques with most of the common datasets to date [13]. All of the techniques have been abandoned in order to assess overall success and ensure that results are repeatable. This is still a problem with our paper, and we're working hard to strengthen it in the future. In the future, we'd like to suggest an attack detection model that examines deep studying processes in order to boost the overall performance of low-frequency assaults [1]. Later on, as IDS techniques are employed to complex and changing network settings, such as Cloud Computing, a variety of problems can be based on them.

**References**

1. KDD Cup Data. Accessed: 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
2. NSL-KDD Dataset. [Online]. Available: https://www.unb.ca/cic/ datasets/nsl.html [3] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, ‘‘A detailed analysis of the KDD CUP 99 data set,’’ in Proc. 2nd IEEE Symp. Comput. Intell. Secur. Defense Appl. (CISDA), Jul. 2009, pp. 1–6.
3. Tarun Dhar Diwan, Dr. Siddhartha Choubey, Dr.H.S.Hota, “An Investigation and Analysis of Cyber Security Information Systems: Latest Trends and Future Suggestion” publication in International Journal IT in Industry, Vol. 9, No.2, 2021, pp. 477–492, ISSN: ISSN (Print): 2204-0595 ISSN (Online): 2203-1731.
4. Tarun Dhar Diwan, Dr. Siddhartha Choubey, Dr.H.S.Hota, “Multifactor Authentication Methods: A Framework for Their Selection and Comparison” publication in International Journal of Future Generation Communication and Networking Vol. 13, No. 3, (2020), pp. 2522–2538, ISSN: 2233-7857 (Web of Science).
5. Tarun Dhar Diwan, Dr. Siddhartha Choubey, Dr.H.S.Hota, “A Novel Hybrid Approach for Cyber Security in IoT network Using Deep Learning Techniques” publication in International Journal of Advanced Science and Technology ISSN:2394-5125, ISSN: 2005-4238 (Scopus indexed Journal).
6. Tarun Dhar Diwan, Dr. Siddhartha Choubey, Dr.H.S.Hota, entitled “Development of Real Time Automated Security System for Internet of Things (IoT)” publication in International Journal of Advanced Science and Technology Vol. 29, No. 6s, (2020), pp. 4180 – 4195, ISSN: 2005-4238 (Scopus indexed Journal).
7. Tarun Dhar Diwan, Dr. Siddhartha Choubey, Dr.H.S.Hota, “A Proposed Security Framework for Internet of Things: An Overview” presented in international Conference held on 20-22 December,2019, MTMI, Inc. USA in Collaboration with at amity Institute of Higher Education, Mauritius.
8. Tarun Dhar Diwan, Dr. Siddhartha Choubey, Dr.H.S.Hota, “Control of Public Services for Public Safety through Cloud Computing Environment” presented in international Conference held on 04-05 January,2020, Organized by Atal Bihari Vajpayee University, Bilaspur in association with MTMI, USA and sponsored by CGCOST, Raipur (C.G), India.
9. Tarun Dhar Diwan, Dr.H.S.Hota, Dr. Siddhartha Choubey “A Study on Security and Data Privacy issues of IoT based Application in Modern Society” presented in international Conference held on 04-05 January,2020, Organized by Atal Bihari Vajpayee University, Bilaspur in association with MTMI, USA and sponsored by CGCOST, Raipur (C.G), India.
10. N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: bot-IoT dataset," arXiv preprint arXiv:1811.00701, 2018

11. Kreibich C, Crowcroft J (2004) Honeycomb: creating intrusion detection signatures using honeypots. SIGCOMM ComputCommun Rev 34(1):51–56
12. Kshetri N, Voas J (2017) Hacking power grids: a current problem. Computer 50(12):91–95
13. P. Laskov, P. Düssel, C. Schäfer, and K. Rieck, "Learning intrusion detection: supervised or unsupervised" in Image analysis and processing – ICIAP 2005: 13th international conference, Cagliari, Italy, September 6–8, 2005. Proceedings, F. Roli and S. Vitulano, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 50–57
14. Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," Expert SystAppl, vol. 39, no. 1, pp. 424–430, 2012/01/01/ 2012
15. Liao H-J, Lin C-HR, Lin Y-C, Tung K-Y (2013b) Intrusion detection system: a comprehensive review. J NetwComputAppl 36(1):16–24
16. H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: a comprehensive review," J NetwComputAppl, vol. 36, no. 1, pp. 16–24, 2013a/01/01/ 2013
17. Lin C, Lin Y-D, Lai Y-C (2011) A hybrid algorithm of backward hashing and automaton tracking for virus scanning. IEEE Trans Comput 60(4):594–601.
18. W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: an intrusion detection system based on combining cluster centers and nearest neighbors," Knowl-Based Syst, vol. 78, no. Supplement C, pp. 13–21, 2015/04/01/ 2015
19. Liu X, Zhu P, Zhang Y, Chen K (2015) A collaborative intrusion detection mechanism against false data injection attack in advanced metering infrastructure. IEEE Transactions on Smart Grid 6(5):2435–2443
20. T. F. Lunt, "Automated audit trail analysis and intrusion detection: a survey," in Proceedings of the 11th National Computer Security Conference, 1988, vol. 353: Baltimore, MD
21. J. Lyngdoh, M. I. Hussain, S. Majaw, and H. K. Kalita, "An intrusion detection method using artificial immune system approach," in International conference on advanced informatics for computing research, 2018, pp. 379–387: Springer
22. McHugh J (2000) Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Trans Inf SystSecur 3(4):262–294
23. C. R. Meiners, J. Patel, E. Norige, E. Torng, and A. X. Liu, "Fast regular expression matching using small TCAMs for network intrusion detection and prevention systems," presented at the Proceedings of the 19th USENIX conference on security, Washington, DC, 2010
24. Meshram A, Haas C (2017) Anomaly detection in industrial networks using machine learning: a roadmap. In: Beyerer J, Niggemann O, Kühnert C (eds) Machine learning for cyber physical systems: selected papers from the international conference ML4CPS 2016. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 65–72
25. Metke AR, Ekl RL (2010) Security Technology for Smart Grid Networks. IEEE Transactions on Smart Grid 1(1):99–107
26. MIT Lincoln Laboratory. (1999, June). DARPA Intrusion Detection Data Sets. Available: https://www.ll.mit.edu/ideval/data/
27. Mitchell R, Chen IR (2015) Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems. IEEE Transactions on Dependable and Secure Computing 12(1):16–30
28. C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in cloud," J NetwComputAppl, vol. 36, no. 1, pp. 42–57, 2013/01/01/ 2013
29. Mohurle S, Patil M (2017) A brief study of wannacry threat: ransomware attack 2017. Int J Adv Res Comput Sci 8(5)

30. S. N. Murray, B. P. Walsh, D. Kelliher, and D. T. J. O'Sullivan, "Multi-variable optimization of thermal energy efficiency retrofitting of buildings using static modelling and genetic algorithms – a case study," Build Environ, vol. 75, no. Supplement C, pp. 98–107, 2014/05/01/ 2014
31. Nourian A, Madnick S (2018) A systems theoretic approach to the security threats in cyber physical systems applied to Stuxnet. IEEE Transactions on Dependable and Secure Computing 15(1):2–13
32. Pasqualetti F, Dörfler F, Bullo F (2013) Attack detection and identification in cyber-physical systems. IEEE Trans Autom Control 58(11):2715–2729
33. Patel, M. Taghavi, K. Bakhtiyari, and J. Celestino Júnior, "An intrusion detection and prevention system in cloud computing: a systematic review," J NetwComputAppl, vol. 36, no. 1, pp. 25–41, 2013/01/01/ 2013
34. Pretorius B, van Niekerk B (2016) Cyber-security for ICS/SCADA: a south African perspective. International Journal of Cyber Warfare and Terrorism (IJCWT) 6(3):1–16
35. T. H. Ptacek and T. N. Newsham, "Insertion, evasion, and denial of service: eluding network intrusion detection," DTIC Document 1998
36. W. Qingtao and S. Zhiqing, "Network anomaly detection using time series analysis," in Joint international conference on autonomic and autonomous systems and international conference on networking and services - (icasisns'05), 2005, pp. 42–42
37. Ge Yan-qiang. Practical Computer Network Security Technology. Bei Jing: Publishing House of China Water Resources and Hydropower, 2010.
38. Marrack Pkappler J. W.How the Immune System Recognizes the Body[J].Scientific American2003269(3), pp. 80-90.
39. Qu Xiao-hongResearch on distributed intrusion detection system based on Protocol analysis Anti-counterfeitingSecurity and Identification in Communication.ASID 2009,3rd International Conference on 20-22 Aug, 2009.
40. Abbasi, J. Wetzels, W. Bokslag, E. Zambon, and S. Etalle, "On emulation-based network intrusion detection systems," in Research in attacks, intrusions and defenses: 17th international symposium, RAID 2014, Gothenburg, Sweden, September 17–19, 2014. Proceedings, A. Stavrou, H. Bos, and G. Portokalidis, Eds. Cham: Springer International Publishing, 2014, pp. 384–404
41. A. Aburomman and M. B. IbneReaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," Appl Soft Comput, vol. 38, pp. 360– 372, 2016/01/01/ 2016
42. Adebowale A, Idowu S, Amarachi AA (2013) Comparative study of selected data mining algorithms used for intrusion detection. International Journal of Soft Computing and Engineering (IJSCE) 3(3):237–241
43. Agrawal S, Agrawal J (2015) Survey on anomaly detection using data mining techniques. Procedia Computer Science 60:708–713