Model performance comparison of CIDDS-001 dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Model | Accuracy | Precision | Recall | F1-score |
| [6] | DT | 99.09 % | 0.99 | 0.99 | 0.99 |
| NB | 60.56 % | 0.47 | 0.61 | 0.48 |
| SVC | 62.89 % | 0.75 | 0.63 | 0.49 |
| Bagging(DT) | 99.08 % | 0.99 | 0.99 | 0.99 |
| Bagging (NB) | 60.57 % | 0.47 | 0.61 | 0.48 |
| Bagging (SVC) | 62.89 % | 0.75 | 0.63 | 0.49 |
| Adaboost (DT) | 99.15 % | 0.99 | 0.99 | 0.99 |
| Adaboost (NB) | 70.74 % | 0.84 | 0.71 | 0.75 |
| Adaboost (SVC) | 62.35 % | 0.39 | 0.62 | 0.48 |
| RF | 99.14 % | 0.99 | 0.99 | 0.99 |
| MajoritytVoting | 63.44 % | 0.76 | 0.63 | 0.50 |
| Proposed | KNN | 99.37% | 99% | 99% | 99% |
| LR | 60.58% | 35% | 45% | 38% |
| DT | 98.86% | 99% | 99% | 99% |
| NB | 68.15% | 34% | 50% | 41% |
| RF | 99.95% | 100% | 100% | 100% |
| SVM | 91.28% | 89% | 93% | 90% |

Model performance comparison of UNSW-15NB dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Model | Accuracy | Precision | Recall | F1-score |
| [9] | KNN | 100% | 100% | 100% | N/A |
| RF | 100% | 100% | 100% | N/A |
| NB | 95.3505% | 95.50% | 95.40% | N/A |
| [10] | DT | 85.56% | N/A | N/A | N/A |
| LR | 83.15% | N/A | N/A | N/A |
| NB | 82.07% | N/A | N/A | N/A |
| ANN | 81.84% | N/A | N/A | N/A |
| EM clustering | 78.47% | N/A | N/A | N/A |
| Voting-CMN | 89.29% | N/A | N/A | N/A |
| V-NKDE | 98.09% | N/A | N/A | N/A |
| [7] | SVM | 84.32% | N/A | N/A | N/A |
| DT | 94.43% | N/A | N/A | N/A |
| NB | 71.63% | N/A | N/A | N/A |
| ANN | 63.97% | N/A | N/A | N/A |
| USML | 94.78% | N/A | N/A | N/A |
| proposed | KNN | 90.89% | 89% | 93% | 90% |
| LR | 89.82% | 88% | 92% | 89% |
| DT | 100% | 100% | 100% | 100% |
| NB | 74.34% | 76% | 79% | 74% |
| RF | 100% | 100% | 100% | 100% |
| SVM | 89.58% | 87% | 91% | 89% |

Model performance comparison of CICIDS2017dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Model | Accuracy | Precision | Recall | F1-score |
| [10] | XGBoost-IDS | 91.36% | N/A | 97.4% | N/A |
| Voting-RKM | 97.77% | 99.83% | 93.13% | 96.36% |
| Enemble CNN | 99.45% | 99.57% | 99.64% | 99.61% |
| V-NKDE | 99.67% | 99.70% | 99.70% | 99.70% |
| [11] | RF | 97.41% | 99.89% | 91.95% | 95.75% |
| KNN | 95.57% | 99.82% | 92.50% | 96.02% |
| MLP | 95.57% | 99.06% | 93.25% | 96.07% |
| Voting-RKM | 97.77% | 99.83% | 93.13% | 96.36% |
| proposed | KNN | 98.82% | 99% | 99% | 99% |
| LR | 74.82% | 76% | 75% | 75% |
| DT | 89.10% | 88% | 89% | 88% |
| NB | 73% | 91% | 73% | 80% |
| RF | 99.93% | 100% | 100% | 100% |

Model performance comparison of NSL-KDD dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Model | Accuracy | Precision | Recall | F1-score |
| [10] | OS-ELM | 98.69% | 98.55% | 98.93% | 98.74% |
| Voting-CKM | 99.68% | 99.74% | 99.57% | 99.66% |
| SMO | 99.40% | N/A | N/A | 95 |
| J48 | 99.75% | N/A | N/A | N/A |
| NB | 95.87% | N/A | N/A | N/A |
| V-NKDE | 99.77% | 99.80% | 99.80% | 99.80% |
| [11] | CART | 98.49% | 98.45% | 98.37% | 92.30% |
| KNN | 99.59% | 99.62% | 99.56% | 99.51% |
| MLP | 99.65% | 99.68% | 99.65% | 99.58% |
| Voting-CKM | 99.68% | 99.74% | 99.66% | 99.57% |
| proposed | KNN | 76.55% | 81% | 79% | 76% |
| LR | 74.17% | 77% | 76% | 74% |
| DT | 77.66% | 77% | 78 | 77% |
| NB | 78.26% | 82% | 80% | 78% |
| RF | 79.62% | 82% | 81% | 80% |

Feature analysis

CIDDS-001 Dataset

The CIDDS-001 is a labelled flow-based dataset and it developed for the evaluation purpose of Anomaly-based Network Intrusion Detection System (NIDS) [12]. CIDDS-001 dataset consists of unidirectional NetFlow data. It consists of traffic data from OpenStack environment having internal servers like sever backup, mail, file, and web and External Servers External Server like file synchronization and web server which is deployed on the internet to capture real-time and up-to-date traffic from the internet. Dataset have three logs files (attack logs, client configuration, traffic logs) and both sever have 4 week captured traffic data. The dataset consists of 14 attributes, where 1-10 are the netFlow default features and attribute 11 – 14 are features added during the labeling process. Four type of attack captured in the dataset (suspicious, attacker, unknown, and victim). Table 1 provides a description for CIDDS-001 dataset features.

|  |  |  |
| --- | --- | --- |
| **No** | **Feature Name** | **Feature Description** |
| 1 | Src IP | IP Address of the source node. |
| 2 | Src\_Port | Port of the source node. |
| 3 | Dest\_IP | IP Address of the destination node. |
| 4 | Dest\_Port | Port of the destination node. |
| 5 | Proto | Transport Protocol (e.g. ICMP, TCP, or UDP). |
| 6 | Date\_first\_seen | Start time flow first seen. |
| 7 | Duration | Flow duration. |
| 8 | Bytes | Number of transmitted bytes. |
| 9 | Packets | Number of transmitted packets. |
| 10 | Flags | OR concatenation of all TCP Flags. |
| 11 | Class | Class label (Normal, Attacker, Victim, Suspicious, and Unknown). |
| 12 | AttackDescription | Provides additional information about the set attack parameters (e.g. the number of attempted password guesses for SSH-Brute-Force attacks). |
| 13 | AttackType | Types of attack (portScan, dos, bruteForce, PingScan). |
| 14 | AttackID | Unique Attack id. Allows attacks which belong to the same class carry the same attack id. |

Table 1. CIDDS-001 dataset features

For this work I used CIDDS-001- external-week1.csv datasets. Here total 172838 instances in the dataset and 70% data use for training and 30% use for test. There type of attack flow in there (Normal, Suspicious, Unknown). Table 2 show, Count of individual attack classes in the dataset.

|  |  |
| --- | --- |
| Label | Number |
| Normal | 49606 |
| Suspicious | 107344 |
| Unknown | 15888 |

Table 2. Count of individual attack classes in the dataset.

**CICIDS2017 dataset**

The CICIDS-2017 dataset was developed to meet the scarcity of realtime network traffic datasets [13]. It contains up-to-date network attacks but also it has all type of real-world attacks. This dataset consists of labelled network flows and full packet payloads in PCAP format [14]. CICIDS2017 dataset consists of eight traffic monitoring sessions all are separated by CSV file that represent the profile of the network traffic for five days. This file contains normal traffic and anomaly traffic. Normal traffic defined as ‘Benign’ traffic and anomaly traffic called as ‘Attacks’ traffic. This dataset has five days traffic data, Thursday and Friday working hour afternoon and also Friday morning data are binary classification. Tuesday, Wednesday, and Thursday morning data for designing a multiclass detection model. The files containing of CICIDS-2017 dataset are displayed in Table 3.

|  |  |  |
| --- | --- | --- |
| **Name of Files** | **Attacks found** | **Flow count** |
| Monday-WorkingHours.pcap\_ISCX.csv | No Attack | 529918 |
| Tuesday-WorkingHours.pcap\_ISCX.csv | Benign, FTP-Patator, SSH-Patator | 445909 |
| Wednesday-workingHours.pcap\_ISCX.csv | Benign, DoS GoldenEye,  DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed | 692703 |
| Thursday-WorkingHours-MorningWebAttacks.pcap\_ ISCX.csv | Benign, Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS | 170366 |
| Thursday-WorkingHours-AfternoonInfilteration.pcap\_ ISCX.csv | Benign, Infiltration | 288602 |
| Friday-WorkingHours-Morning.pcap\_ISCX.csv | Benign, Bot | 191033 |
| Friday-WorkingHours-AfternoonPortScan.pcap\_ISCX.csv | Benign, PortScan | 225745 |
| Friday-WorkingHours-AfternoonDDos.pcap\_ISCX.csv | Benign, DDoS | 286467 |

Table 3. Descriptions of files containing CICIDS-2017 dataset

The dataset has 79 number of features and whole shape of a dataset that contains 2830743 instances. Also containing 15 class labels 1 is normal traffic and 14 is attack traffic. Table 4 show the characteristics of the CICIDS2017dataset.

|  |  |
| --- | --- |
| Year of release | 2017 |
| Total number of distinct instances | 2830743 |
| Number of features | 79 |
| Number of classes | 15 |

Table4.The characteristics of the CICIDS2017dataset.

**UNSW-15NB dataset**

The UNSW-NB 15 dataset was generated in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) by the IXIA Storm tool to extract a hybrid of modern normal and modern attack behaviors [15].The dataset have 49 attribute with class label and contains 2, 540,044labelled instances. The features are categorized into seven groups that include flow features, basic features, content features, time features, general purpose, connection features, and class label [16]. Table 5 given in a detailed description about the features.

|  |  |  |
| --- | --- | --- |
| S.No  . | Type of  attributes | Name of attributes |
| 1 | Flow | Script, Sport, Dstip. Dsport, Proto |
| 2 | Basic | State, Dur, Sbytes, Dbytes, Sttl, Dttl, Sloss, Dloss, Service, Sload, Dload, Spkts, Dpkts |
| 3 | Content | Swin, Dwin, Stepb, Dtcpb, Smeansz, Dmeansz, trans\_depth, res\_bdy\_len |
| 4 | Time | Sjit, Djit, Stime , Ltime, Sintpkt, Dintpkt, Tcprtt, Synack, Ackdat |
| 5 | General  Purpose | is\_sm\_ips\_ports, ct\_state\_ttl, ct\_flw\_http\_mthd,is\_ftp\_login, ct\_ftp\_cmd |
| 6 | Connection | ct\_srv\_src , ct\_srv\_dst , ct\_dst\_ltm, ct\_src\_ ltm ,ct\_src\_dport\_ltm,ct\_dst\_sport\_ltm,ct\_dst\_src\_ltm |
| 7 | Labelled | attack\_cat, Label |

Table 5: Description of the attributes of UNSW-NB15 dataset

In the dataset feature being some multiple types of value, some being nominal, some being numeric like Integer, Binary and Float and some taking on timestamp values. Table 6 show that,

|  |  |  |
| --- | --- | --- |
| **No.** | **Feature Type** | **Count** |
| 1 | Nominal | 6 |
| 2 | Integer | 28 |
| 3 | Binary | 3 |
| 4 | Float | 10 |
| 5 | Timestamp | 2 |

Table 6. Features type of UNSW-NB15 dataset

UNSW-NB15 dataset contain nine type of attacks and one represent normal attack in the dataset. The attacks are categorized as Fuzzers, Reconnaissance, Shellcode, Analysis, Backdoors, DoS, Exploits, Generic, and Worms. Total 2, 540,044 number of instances take on the dataset. which are stored in the four CSV files. The number of records in the training set is 175,341 records and the testing set is 82,332 records from the different types, attack and normal. Table 7 , showed Details of instances and Table 8 showed, type of attacks and the number in UNSW-NB 15 dataset.

|  |  |
| --- | --- |
| **Name** | **Count** |
| Total Number of events | 2540044 |
| Normal | 2218761 |
| Attacks | 321283 |

Table 7. Details of instances in UNSW-NB 15 dataset

|  |  |  |
| --- | --- | --- |
| **Type** | **Whole Dataset** | **Training Dataset** |
| Normal | 2218761 | 56000 |
| Fuzzers | 24246 | 18184 |
| Analysis | 2677 | 2000 |
| Backdoors | 2329 | 1746 |
| DOS | 16353 | 12264 |
| Exploits | 44525 | 33393 |
| Generic | 215481 | 40000 |
| Reconnaissance | 13987 | 10491 |
| ShellCode | 1511 | 1133 |
| Worms | 174 | 130 |

Table 8.Type of attacks in UNSW-NB 15 dataset

**NSL-KDD dataset**

The NSL-KDD dataset from Canadian Institute of Cybersecurity it develops by detect flow-based instance form network using machine learning [17]. It’s an upgraded version of KDD99 dataset. KDD99 dataset have so many issues one of most important issues is lots of duplicate and redundant instance in the dataset. But NSL-KDD overcome those problem and data packets represent current real network-based IDS. Dataset have two parts one is training dataset and another one is test dataset. Dataset have 42 features and instance classifies as a normal or attack. Table 9, show NSL\_KDD dataset features. Four type of group attack classes instance in dataset and they are Dos, Probe, R2L, and U2R. The training dataset consists of 23 classes and the testing dataset consists of 38 classes that include 21 attacks from training dataset. Table 10 show NSL-KDD attack types a with there classes.

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Features** | **No** | **Features** |
| 1 | duration lenght | 22 | is\_guest\_login |
| 2 | protocol\_type | 23 | count |
| 3 | service | 24 | srv\_count |
| 4 | flag | 25 | serror\_rate |
| 5 | src\_bytes | 26 | srv\_serror\_rate |
| 6 | dst\_bytes | 27 | rerror\_rate |
| 7 | land | 28 | srv\_rerror\_rate |
| 8 | wrong\_fragment | 29 | same\_srv\_rate |
| 9 | urgent | 30 | diff\_srv\_rate |
| 10 | hot | 31 | srv\_diff\_host\_rate |
| 11 | num\_failed\_logins | 32 | dst\_host\_count |
| 12 | logged\_in | 33 | dst\_host\_srv\_count |
| 13 | lnum\_compromised | 34 | dst\_host\_same\_srv\_rate |
| 14 | lroot\_shell | 35 | dst\_host\_diffsrv\_rate |
| 15 | lsu\_attempted | 36 | dst\_host\_same\_src\_port\_rate |
| 16 | lnum\_root | 37 | dst\_host\_srv\_diff\_host\_rate |
| 17 | lnum\_file\_creations | 38 | dst\_host\_serror\_rate |
| 18 | lnum\_shells | 39 | dst\_host\_srv\_serror\_rate |
| 19 | lnum\_access\_files | 40 | dst\_host\_rerror\_rate |
| 20 | lnum\_outbound\_cmds | 41 | dst\_host\_srv\_rerror\_rate |
| 21 | is\_hot\_login | 42 | Class |

Table 9. NSL\_KDD dataset features

|  |  |
| --- | --- |
| Attack | Attack classes |
| Normal | normal |
| DOS | smurf, teardrop, neptune, back, pod, land |
| Probe | ipsweep, portsweep, nmap, satan |
| R2L | phf, guess\_passwd, spy, warezmaster, ftp\_write, warezclient, imap, multihop |
| U2R | buffe\_overflow, loadmodule, perl, rootkit |

Table 10. NSL-KDD attack types

NSL-KDD the training dataset consists of 125,973 instances and testing dataset consists of 22,544 instances. Table 11,12 showed count labels based on individual attack classes is indicated training and test dataset.

|  |  |
| --- | --- |
| **Class** | **Count Instances** |
| Normal | 67343 |
| DOS | 45927 |
| Probe | 11656 |
| R2L | 995 |
| U2R | 52 |
| Total | 125973 |

Table 11. NSL\_KDD training dataset Count of individual attack classes

|  |  |
| --- | --- |
| **Class** | **Count Instances** |
| Normal | 9711 |
| DOS | 7458 |
| Probe | 2421 |
| R2L | 2754 |
| U2R | 200 |
| Total | 22544 |

Table 12. NSL\_KDD test dataset Count of individual attack classes.