Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549507

# COMPARATIVE ANALYSIS OF RESAMPLING TECHNIQUES ON IMBALANCED CIE-CICIDS2018 DATASET FOR DOS ATTACK DETECTION

**Supriya Dicholkar, Jagannath Nirmal**
Department of Electronics Engineering
K J Somaiya College of Engineering,Vidyavihar,India-400077,
Email ID: supriya.ganu@gmail.com, jhnirmal@somaiya.edu

**Abstract—** The Internet of Things (IoT) is a largely emerging area having applications in almost all sectors but a threat to the security of the IoT network is the main hurdle in the growth of IoT networks. For attack detection, standard IoT datasets are used. These datasets are highly imbalanced with major benign traffic and very little attack traffic. To deal with the imbalanced dataset in this paper, different resampling techniques such as Undersampling, Oversampling, and hybrid sampling are applied to the CIE-CICIDS2018 dataset. After resampling, the artificial neural network is applied for attack detection on this resampled dataset. As the dataset is imbalanced, for evaluation of the performance along with accuracy, precision, recall, and the F1 score are parameters used. Random Undersampling is giving the best result among all resampling techniques but a lot of data loss occurred in Random Undersampling. Edited Nearest neighbors is giving better results than all other techniques except Random Undersampling without losing the majority of data samples.
**Keywords:** class imbalance,CICIDS2018, IDS, resampling,

## I. Introduction

The Internet of Things (IoT) is nothing but any object in the surroundings such as sensors, software, and other objects connected to the Internet. Kelvin Ashton 1999 coined the term 'IoT' first time when Radio Frequency Identification (RFID) was used with computers for the integration of people, processes, and technologies. Nowadays, many applications such as smart homes, smart grids, smart parking, smart agriculture, etc. are evolving with the growth of IoT and sensor technology. The growth of IoT is tremendous in this decade and as per Gartner's report by 2025, more than 64 billion IoT devices worldwide will be used (Gartner Research).

With such vast growth of IoT, there are certain challenges faced by users. IoT devices are low-constrained devices having less battery life, less memory, and limited storage. Heterogeneity in IoT devices and the security of IoT devices are some of the common hurdles in IoT device expansion. A major challenge is maintaining confidentiality, Integrity, and Availability of data transferred between IoT devices (Dicholkar and Sekhar 2020). Different approaches have been used till now for the detection and mitigation of attacks on IoT devices such as the use of honeypots, firewalls, etc but the use of machine learning and deep learning for attack detection is a very useful and efficient approach as IoT devices are creating enormous data every day. These datasets are imbalanced datasets having a majority of benign traffic and very little attack traffic.

In this paper, different resampling techniques are applied to the CSE-CIC-IDS2018 dataset for balancing the dataset, and Artificial Neural Network is applied for attack detection. A comparison of different resampling results is discussed in section 4.
The key contributions of this paper are as below:

[1] This is the first paper comparing all undersampling, oversampling, and hybrid techniques on the latest CSE-CIC-IDS2018 dataset.
[2] For comparing the performance of different resampling techniques, along with accuracy, f1 score, precision, and recall are also used as the dataset is an imbalanced dataset.
[3] Performance comparison of the proposed model with the state-of-the-art systems is carried out.

The organization of the article is as follows: Section 1 contains an introduction to IoT security. Section 2 contains work related to a class imbalance of the dataset and different techniques for resolving this problem. Section 3 contains a model demonstration used for handling an Imbalanced dataset. Section 4 contains result discussion obtained from different undersampling, oversampling, and hybrid techniques. Section 5 contains the conclusion and future work.

## 2. Related work

A lot of research work is going on based on the detection of attacks in IoT networks using machine learning and deep learning. The accuracy of the designed model is dependent on the dataset used for analysis. The researcher used some standard datasets such as Kdd-Cup99, NSL-KDD, CICIDS2017, and CSE-CIC-IDS2018. The main problem with these datasets is all datasets are highly imbalanced having very large benign traffic and very fewer attacks traffic such as SQL injection, infiltration, etc. (Karatas et al. 2020). Results obtained from such databases are giving very good accuracy but results are biased towards the majority of benign traffic. For an imbalanced dataset, along with accuracy, F1 score and recall are to be used for the detection of minority class attack traffic properly. For getting better results from these datasets, solving the class imbalance problem is very much essential.

In 2016, Ajinkya et.al. applied different undersampling such as Random Undersampling, Near Miss-1, Near Miss-2, Tomek, Condensed Nearest Neighbours, Edited Nearest Neighbours, etc., oversampling such as Random oversampling, SMOTE, Borderline SMOTE, etc. and hybrid techniques such as SMOTE ENN and SMOTE Tomek to the synthetic dataset. Along with accuracy, precision, and recall are parameters considered for comparison as the dataset is an imbalanced dataset. Among all techniques, SMOTE + ENN is giving the best results for this dataset (More 2016). In 2020, Young et.al. compared three oversampling methods SMOTE, borderline SMOTE, and ADASYN on synthetic records of 299 heart patients. SMOTE is giving the best result with F1 score of 0.63 as compared to borderline SMOTE with F1 Score of 0.6 and ADASYN with F1 Score of 0.62 (Kim et al. 2020).

In 2020, hongopo et.al. used a hybrid method in which undersampling with Gaussian Mixture Model is carried out on CICIDS2017 and UNSW-NB15 datasets. and oversampling is done with SMOTE. On a resampled dataset, Convolutional Neural Network (CNN), Random Forest(RF), and Multi-Layer perceptron (MLP) are applied. CNN is giving the best results with an accuracy of 98.82 % and F1 score of 95.53% whereas MLP and RF are giving accuracies of 98.74 % and 98.68% respectively and F1 scores of 95.25 and 95.03 respectively for MLP and RF for the UNSW-NB15 dataset. SGM is also applied to the CICIDS2017 dataset and the accuracy is 99.85% but the main drawback of this algorithm is that the time required for training and testing is very large (Zhang et al. 2020).

In 2021. Shikha et. al. applied random undersampling, random oversampling, random undersampling and random oversampling, random undersampling with SMOTE, and random undersampling with Adaptive Synthetic sampling applied on KDD99, UNSW-NB15, UNSW-NB17, and UNSW-NB18. After resampling, Artificial Neural Network(ANN) is applied. Random undersampling is giving almost better results with 93.12% macro precision and 90.37% macro recall 87.19% F1 score among all techniques with the least time of 834 seconds (Bagui 2021).

## 3. Datasets

For designing a network intrusion detection system (NIDS), datasets with different class samples such as benign and different attack traffic such as DoS, SQL injection, infiltration, etc. should be in an adequate proportion. Most of the datasets available are Imbalanced with a majority of benign traffic and very less attack traffic. Researchers can use standard datasets or generate their own datasets with benign and attack traffic. Different eight benchmark datasets used for attack detection with their traffic are KDD Cup99, CAIDA UCS 2007, NSL-KDD, ADFA-LD/WD CIC-IDS2017, and CIE-CIC-IDS2018 (Tavallaee et al.2009 ; Panigrahi et al.2018 ; Kharaisat et al. 2019 ) and Dataset with their properties are compared in table 1 for finalizing the dataset for our work. It can be seen from Table 1 that CIE-CII-IDS2018 is the latest dataset with all the required features.

Table 1. Different dataset Description

| Sr. No. | Features | KDD CUP99 | CAIDA UCS 2007 | NSL-KDD | ADFA-LD/WD | CIC-IDS2017 | CIE-CIC-IDS2018 |
|---|---|---|---|---|---|---|---|
| 1 | Complete Network Config. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 2 | Complete Traffic | × | ✓ | × | ✓ | ✓ | ✓ |
| 3 | Labelled Dataset | ✓ | × | ✓ | ✓ | ✓ | ✓ |
| 4 | Complete Interaction | ✓ | × | ✓ | ✓ | ✓ | ✓ |
| 5 | Complete Capture | ✓ | × | ✓ | ✓ | ✓ | ✓ |
| 6 | Protocols not covered | https | https, SSH, FTP, email | https | https | All covered | All covered |
| 7 | Attack Diversity | R2L, U2L, DoS, Probing attack | DDoS | R2L, U2L, DoS, Probing attack | Hydra-FTP, Hydra-SSH Adduser, Java meterpreter, Webshell | Bot, Brute force, DoS, SQL injection, infiltration | Bot, Brute force, DoS, SQL injection, infiltration |
| 8 | Heterogeneity | × | × | × | | ✓ | ✓ |
| 9 | Feature Set | ✓ | × | ✓ | × | ✓ | ✓ |
| 10 | Metadata | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

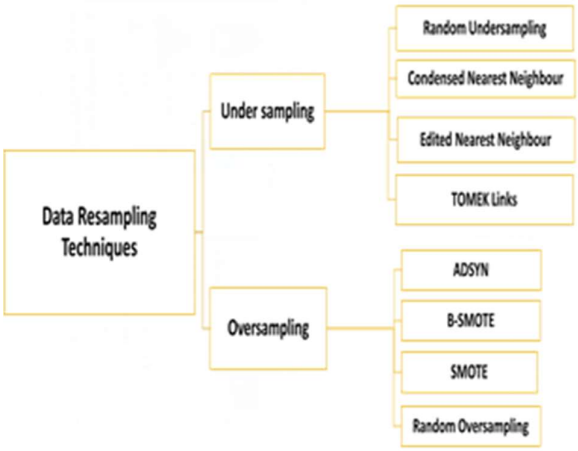
## 4. Class Imbalance

The class imbalance problem in the dataset is nothing but uneven distribution of class samples i.e. suppose the dataset is having 90% samples from the majority class (normal traffic) and the remaining 10\% samples from the minority class (attack traffic) (Koroniotis et al. 2019). For such an imbalanced dataset, pre-processing is required before applying machine learning techniques. For handling class imbalance in the dataset, the researcher used either data-level solutions such as resampling or algorithmic approaches such as cost-sensitive classification (Puri and Gupta 2019). D. Devi explained different pure undersampling methods such as ENN, Tomek, CNN, etc., and hybrid techniques which are a combination of pure undersampling and clustering or ensemble techniques for handling class imbalance (Devi et al. 2020). It is concluded in this paper that a hybrid method evolved from two or more undersampling techniques is more effective than a single technique.

### 4.1 Resampling Techniques

In Data Resampling techniques, two main techniques available are undersampling and oversampling (Amin et al. 2015 and Afreen et al. 2022). In undersampling, data samples of the majority class will be reduced and tried to match with minority samples so the total samples of the resampled dataset are less than the original dataset. In oversampling, data samples of minority classes will be increased and tried to match with the majority data samples. In oversampling, the overall size of the resampled dataset is more than the original dataset (Amin et al. 2015). Different undersampling and oversampling techniques are shown in Figure 1. Different types of undersampling techniques used in this research work are Random Undersampling, Condensed Nearest Neighbours, Edited Nearest Neighbours, and Tomek links. Different types of oversampling techniques used in this research work are Adaptive Synthetic Sampling (AdaSyn), Borderline Synthetic Minority Oversampling Technique(B-SMOTE), and Random Oversampling.

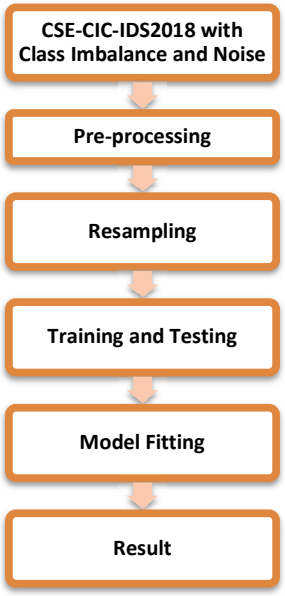
**Figure 1. Different Resampling Techniques**




## 5. Methodology

In this paper, performance analysis of different resampling techniques is carried out using the CSE-CIC-IDS2018 dataset. One CSV file named 'Thursday-15-02-2018_TrafficForML_CICFlowMeter' is used in which only two types of samples namely benign and DoS are present. A flowchart for handling an Imbalanced dataset is shown in Figure. 2.

**Figure. 2 Flowchart of handling Imbalanced dataset**




Before applying resampling techniques, pre-processing of the dataset is carried out such as removing infinity values and normalization done by applying a min-max scaler to scale down all values between zero and one. After pre-processing in stage one, Artificial Neural Network(ANN) is developed without resampling with an input layer, two hidden layers, and an output layer. At the input layer of ANN, 78 neurons are used for 78 input features. Rectified Linear Unit (ReLU) is the activation function used which will convert all negative data to zero and keep positive data as it is. Two hidden layers each having 1000 neurons and ReLU as activation function is used. At the output layer, only one neuron with a sigmoid activation function is used. Adam optimizer is used with 5 epochs as it is more efficient with fewer

memory requirements. Precision, recall, accuracy, and F1 score for ANN without resampling are calculated

In stage two different resampling techniques such as undersampling (Random Undersampling, Tomek, Edited nearest neighbors, Ensemble undersampling), oversampling (Random Oversampling, SMOTE, ADASYN), and hybrid techniques (SMOTE Tomek) are applied to the dataset (Park et al. 2021 and Ahmad et al. 2018). After applying different resampling, the ANN model with previous parameters is developed on resampled data. Precision, recall, accuracy, and F1 score for ANN with resampling are calculated and compared (Ahmad et al. 2021).

### 5.1 CSE-CIC-IDS2018 Dataset

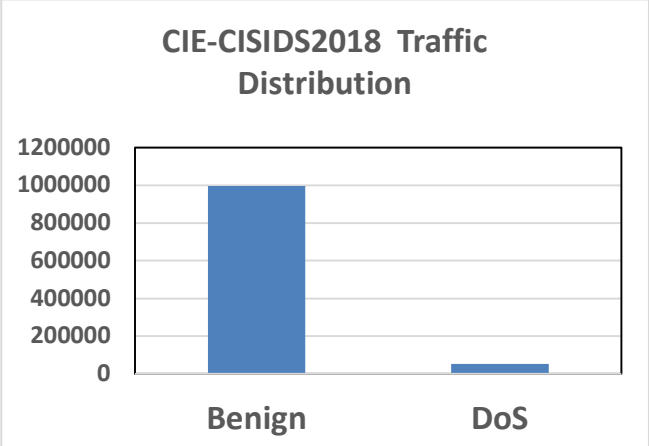
The dataset used for attack detection is CSE-CIC-IDS2018 which has ten CSV files. Different classes present in these CSV are Benign, DoS, Bot, Brute Force, SQL Injection, and Infiltration. The total data for ten CSV files is 45,25,399. For a review of different resampling techniques, one file is taken named 'Thursday-15-02 2018_TrafficForML_CICFlowMeter' which consists of a total of 10,40,548. Class distribution for this CSV is shown in Table 2.

**Table 2. Class Imbalance Distribution in Sample Dataset**

| Total Samples | Class Type | Samples | Imbalance Ratio |
|---|---|---|---|
| 10,48,575 | Benign | 996077 | 94.99 |
| | DoS | 52498 | 5.01 |

From Table 2, it can be seen that this is a clear example of class imbalance in Figure.3.

Figure. 3 Imbalanced CIE-CICDS2018 dataset distribution with DoS attack traffic




Application of simple machine learning or deep learning will give false accuracy for majority samples and very less accuracy for minority samples. The application of resampling techniques will help in the improvement of the performance of the model (Abubakar et al. 2015).

### 6. Result

For an Imbalanced dataset, considering accuracy as an evaluation parameter is not sufficient. A majority of samples in the dataset are benign which will be detected correctly whereas

minority attack traffic will not be detected properly. The detection of attack traffic is the main moto of research work(Manimurugan et al. 2020). In the survey paper, it was shown that accuracy, precision, F1 score, and recall are to be considered while designing IDS with an imbalanced dataset (Tahsien et al. 2020). The confusion matrix for the model will be given in Table 3.

**Table 3. Confusion Matrix**

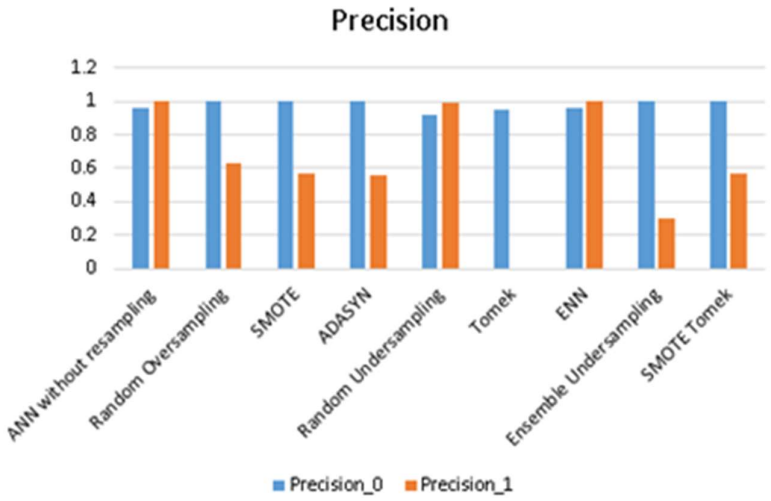
| | | **Predicted Class** | |
|---|---|---|---|
| | | **Attack** | **Normal** |
| **Actual Class** | **Attack** | True Positive(TP) | False Negative(FP) |
| | **Normal** | False Positive(FP) | True Negative(TN) |

Precision is the ratio of correctly predicted attacks to total attacks predicted (Leevy and Khoshgoftaar 2020) and it is given by:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

The results of ANN with and without different resampling techniques are shown in Figure. 4.

Figure 4 Precision of resampling techniques

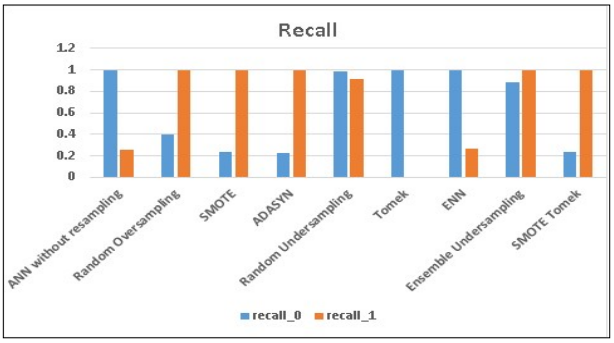



From Fig. 4, it can be observed that the precision value for ENN for benign is 96% and for the attack is 100% whereas the value for Random undersampling for benign is 92% and for the attack is 99%. ENN and Random Undersampling are giving better results as compared to other undersampling and oversampling techniques. Other techniques though are giving good results for benign traffic, precision results for attack traffic are very bad so actual attacks will not be predicted as attacks for other techniques.

Along with precision, recall is another important parameter to be considered for deciding the best method for attack detection. Recall or Detection Rate is the ratio of correctly classified attacks to all actual attacks [20] and it is given by:

$$Recall = Detection\ Rate = \frac{TP}{TP+FN}. \quad (2)$$

Figure. 5 Recall of resampling techniques

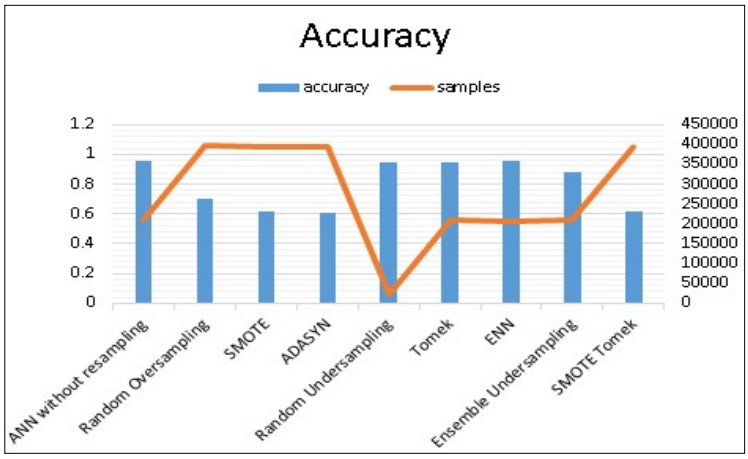



From Figure. 5 it can be seen that the recall value for Random Undersampling for benign is 99% and for the attack is 91% whereas Ensemble undersampling is giving recall value for benign as 88% and for the Attack as 100%. Ensemble undersampling and random Undersampling are giving better results as compared to other techniques. Fewer values of recall indicate some attacks will not be detected as attacks.

Accuracy is the overall accuracy of the attack detection system. Accuracy is the ratio of correctly detected samples (attacks and Benign) to the total number of samples[22] and it is given by:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+F} \quad (3)$$

Figure. 6 Accuracy of resampling techniques



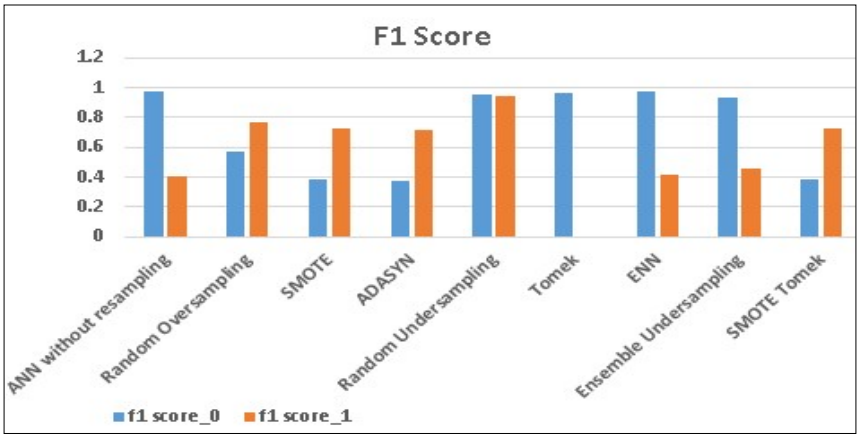

From Figure. 6, it can be seen that the accuracy for Random undersampling is 95%, and for the Edited nearest neighbors is 96%. Random Undersampling and Edited Nearest neighbors are giving better overall accuracy than other techniques. Also, it can be seen that as the number of total samples increases accuracy is reduced. It can be concluded that undersampling techniques are giving better results as compared to oversampling or hybrid techniques.

Along with precision, recall, and accuracy, the F1 score is an important parameter to be considered for evaluation.

The F1 score is the harmonic mean of precision and recall. It is a measure of overall accuracy. For an imbalanced dataset, the F1 score is a very important parameter to be considered. [20] It is given by:

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Rec} \quad (4)$$

Figure. 7 F1 Score of resampling techniques




From Figure. 7, it can be seen that the F1 score for Random undersampling for benign is 96% and for attack, it is 95%. Random Undersampling is giving the best results among all techniques.

Overall comparison of all resampling techniques based on precision, recall, F1 score and Accuracy for Benign as well as attack traffic is shown in Table 4.

Table 4. Performance comparison of resampling techniques

| Sr. No. | Modelling Technique | Model Name | Precision | | Recall | | F1 Score | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| | | | Benign | Attack | Benign | Attack | Benign | Attack | Benign |
| 1 | Oversampling | Random Oversampling | 1 | 0.63 | 0.4 | 1 | 0.57 | 0.77 | 0.7 |
| 2 | | SMOTE | 0.58 | 1 | 1 | 0.27 | 0.73 | 0.42 | 0.63 |
| 3 | | Adasyn | 1 | 0.56 | 0.23 | 1 | 0.37 | 0.72 | 0.61 |
| 4 | Undersampling | **Random Undersampling** | **0.92** | **0.99** | **0.99** | **0.91** | **0.96** | **0.95** | **0.95** |
| 5 | | Tomek | 0.95 | 0 | 1 | 0 | 0.97 | 0 | 0.95 |
| 6 | | Edited Nearest Neighbour | 0.96 | 1 | 1 | 0.27 | 0.98 | 0.42 | 0.96 |
| 7 | | Ensemble Undersampling | 1 | 0.3 | 0.88 | 1 | 0.93 | 0.46 | 0.88 |
| 8 | Hybrid Sampling | SMOTE Tomek | 1 | 0.57 | 0.24 | 1 | 0.39 | 0.73 | 0.62 |

It can be clearly seen that all the values for random undersampling are the best as compared with other undersampling, oversampling, and hybrid sampling methods so Random Undersampling is the best method to be used with the CIE-CIS2018 dataset as per the results obtained with the ANN algorithm.

The proposed method results with a random undersampling are also compared state of art systems based on the accuracy, precision, recall, and F1 score. It can be seen that the proposed method- Random Undersampling with ANN gave the best results than the other two state of art systems.

Table 5

Comparison between the Proposed study and state-of-the-art study

| Study | Dataset | Resampling Technique | Classifier | Accuracy (%) | Precision (%) | Recall (%) | F1 score (%) |
|---|---|---|---|---|---|---|---|
| **Proposed** | CIE-CICIDS2018 | Random Undersampling | ANN | 95 | 99 | 99 | 95 |
| Silva, Bruno. et al. | CIE-CICIDS2018 | Random Undersampling | KNN | 85 | 83 | 85 | 82 |
| Bagui, S. et al. | KDD99 | Random Undersampling | ANN | - | 71.083 | 88.41 | 76.13 |

## 7. Conclusion and future work

In this paper, a detailed analysis of different resampling techniques namely Undersampling, Oversampling, and Hybrid techniques is carried out with the help of the CSE-CIC-IDS2018 dataset. The objective of the paper is to find the best method of resampling for dealing with an imbalanced dataset for attack detection.

Random Undersampling gives the best results for all four parameters: accuracy, precision, recall, and f1 score. Also, ENN is giving better results for precision, accuracy, f1 score, and recall as compared to the remaining resampling techniques. Also, it is been observed that when the number of total samples are increasing after applying either oversampling or hybrid sampling, performance parameter values are decreasing as well as time required for training as well as testing increases. Undersampling methods are the best choice for CSE-CIC-IDS2018 dataset resampling. The future scope of this research work is considering all five attacks for detection as well as anomaly detection. Reduction in the time required for attack detection can be carried out by applying different feature selection methods before applying ANN to the pre-processed dataset.

## Data availability statement

The data that support the findings of this study are openly available at https://www.unb.ca/cic/datasets/ids-2018.html.

## REFERENCES

Abubakar AI, Chiroma H, Muaz SA, Ila LB. 2015, A review of the advances in cyber security benchmark datasets for evaluating data-driven based intrusion detection system. Procedia Comput. Sci. 62:221-227.

Ahmad Z, Khan AS, Shiang CW, Abdullah J, Ahmad Z. 2021. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Wiley. 32(1).

Amin A, Rahim F, Ali I, Khan C, Anwar S. 2015. A comparison of two oversampling techniques (SMOTE vs MTDF) for handling class imbalance problem: A case study of customer churn prediction. conference on New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing. 353:215-225.

Arefeen MA, Nimi ST, Rahman MS. 2022. Neural network-based undersampling techniques. IEEE Trans. Syst Man Cybern. 52(2):1111-1120.

Bagui SK, Li. 2021. Resampling imbalanced data for network intrusion detection datasets. J. Big Data. 8.

Basheri AI, Iqbal MJ, Rahim A. 2018. Performance comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for intrusion detection. IEEE Access. 6: 33789-33795.

Devi D, Biswas SK, Purkayastha B. 2020. A review on solution to class imbalance problem: undersampling approaches. IEEE Conference on ComPE: 626-631.

Dicholkar SV, Sekhar D. 2020. Review-IoT security research opportunities. IEEE International Conference on Convergence to Digital World - Quo Vadis. p. 1-4.

Gartner RESEARCH PORTAL

Karatas G, Demir O, Sahingoz OK. 2020. Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. IEEE Access. 8:32150-32162.

Khraisat A, Gondal I, Vamplew P. 2019. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity.2.

Kim YT, Kim DK, Kim H, Kim DJ. 2020. A comparison of Oversampling methods for constructing a prognostic model in the patient with heart failure. IEEE International Conference on ICTC. p. 379-383.

Koroniotis N, Moustafa N, Sitnikova E, Turnbull B. 2019. Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. Future Gener. Comp. Sy. 100:779-796.

Leevy JL, Khoshgoftaar TM. 2020. A survey and analysis of intrusion detection models based on CSE-CIC- IDS2018. J Big Data, 7(104).

Manimurugan S, Al-Mutairi S, Aborokbah MM, Chilamkurti N, Ganesan S, Patan R. 2020. Effective attack detection in Internet of Medical Things smart environment using a Deep Belief Neural Network. IEEE Access. 8:77396-77404.

More A. 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets. ArXiv.1608.06048.

PANIGRAHI R, BORAH S. 2018. A detailed analysis of CICIDS2017 dataset for designing intrusion detection systems. Int. J. Eng. Technol.7: 479-482.

Park S, Park H. 2021. Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic. Computing. 103:401–424.

Puri A, Gupta MK. 2019. Comparative analysis of resampling techniques under Noisy imbalanced datasets. IEEE (ICICT).p.1-5.

Tahsien SM, Karimipour H, Spachos P. 2020. Machine learning based solutions for security of Internet of Things (IoT): a survey. J. Netw. Comput. Appl. 161( 102630).

Tavallaee M, Bagheri E, Lu W, Ghorbani A. 2009. A detailed analysis of the KDD CUP 99 data set. IEEE CISDA.2:1-6.

Zhang H, Huang L, Wu CQA. 2020. An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. Comput. Netw.177(107315).

Silva, Bruno & Silveira, Ricardo & Neto, Manuel & Cortez, Paulo & Gomes, Danielo. (2021). A comparative analysis of undersampling techniques for network intrusion detection systems design. Journal of Communication and Information Systems. 36. 31-43. 10.14209/jcis.2021.3.

Bagui, S., Li, K. Resampling imbalanced data for network intrusion detection datasets. J Big Data 8, 6 (2021). https://doi.org/10.1186/s40537-020-00390-x