IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 5, October, 2023, pp. 1229-1236
Journal homepage: http://iieta.org/journals/isi

Evaluating the Efficacy of Resampling Techniques in Addressing Class Imbalance for Network Intrusion Detection Systems Using Support Vector Machines

Swarnalatha Kudithipudi[1], Nirmalajyothi Narisetty[2*], Gangadhara Rao Kancherla[1], Basaveswararao Bobba[1]

[1] Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur 522510, Andhra Pradesh, India
[2] Department of CSE-AIML & IOT, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad 500090, Telangana, India

Corresponding Author Email: nirmala.narisetty@gmail.com

https://doi.org/10.18280/isi.280511

Received: 12 June 2023
Revised: 2 September 2023
Accepted: 13 October 2023
Available online: 31 October 2023

Keywords:
class imbalance, Network Intrusion Detection Systems, resampling methods, cloud computing, Canadian Institute for Cyber Security Intrusion Detection dataset-2017, Support Vector Machine (SVM) classifier

ABSTRACT

The objective of this study was to assess the performance of various resampling strategies aimed at mitigating the class imbalance problem in Network Intrusion Detection Systems (NIDS) using machine learning models and imbalanced benchmark datasets. Due to this class imbalance problem, detection of known or unknown attacks in NIDS often results in suboptimal performance. Resampling methods, statistically designed to generate synthetic samples from existing datasets, were employed to rebalance class labels and train the machine learning models. The Support Vector Machine (SVM), a robust supervised classifier, was utilized to classify data by identifying the optimal decision boundary that maximally separates different classes. In this context, efforts were made to enhance the effectiveness of these resampling techniques and consider the potential benefits of hybrid models. No resampling (NR), Synthetic Minority Over-sampling Technique (SMOTE), Random Under Sampling (RUS), Random Under Sampling and Random Over Sampling (RUS+ROS), and Random Under Sampling and SMOTE (RUS+SMOTE) were evaluated. The SVM classifier with Radial Basis Function (RBF) was employed, validated against the imbalanced benchmark dataset CICIDS-2017 (Canadian Institute for Cyber Security Intrusion Detection dataset-2017), to assess the effectiveness of these methods using performance metrics such as Accuracy, Precision, Recall, F1 score, and wall time. The proposed method achieved a remarkable accuracy of 99.63% in intrusion detection, demonstrating impressive results when compared to state-of-the-art methods for detecting network attacks on imbalanced datasets. The findings from this research provide valuable insights into the potential of various resampling methods in tackling class imbalance problems in NIDS.

1. INTRODUCTION

Cloud technology, which provides on-demand network access to computing resources across various domains worldwide, has formed the backbone of the current digital era. Its primary objective is to offer customers an array of services under a pay-as-you-use model, requiring minimal supervision. Two of the most prevalent fields where cloud technology has been adopted are global communication and networking, where an enormous number of end-users and devices are connected to the cloud in cyberspace to access an array of amenities. The benefits of cloud computing, either in terms of service delivery or economic efficiency, are numerous. However, despite these advantages, the rapid advancement in communication technology has introduced numerous security and privacy challenges, such as maintaining confidentiality, integrity, and availability, leading to the emergence of a new class of cyber-attacks [1].

Traditional security measures, including firewalls and encryption of sensitive data, are widely deployed. However, they are increasingly considered outdated for organizations requiring robust security measures, such as government entities and military bases, largely due to the difficulties in human configuration and the extended timeframes required to develop advanced solutions for these attacks [2].

Among these emerging attacks, Distributed Denial of Service (DDoS) attacks hold significant importance due to their potential to critically impact cloud servers and consumers. DDoS attacks purposefully target websites, storage, cloud-hosted applications, and network setups, absorbing all available bandwidth and disrupting access for legitimate users and partners. This can tarnish the reputation of companies by reducing their productivity and affecting their profitability. These disruptive activities are frequently conducted through innovative network penetration methods by unauthorized users.

Although the cloud's elasticity principle ensures that service delivery remains uninterrupted, the service provider's bill is inevitably increased to maintain the Quality of Service (QoS) as per the Service Level Agreement (SLA). Therefore, DDoS attacks often lead to Economic Denial of Service (EDoS) attacks [3]. As such, it is crucial to mitigate these types of attacks in the cloud before the billing mechanism commences for the service provider. To combat these attacks, numerous approaches for intrusion prevention and detection systems have been developed and documented in the literature. However, to ensure the cloud's survival, there is a growing need for more robust mechanisms to defend against

increasingly creative and sophisticated attacks. A significant portion of cloud attacks are network-based, and with the volume of data increasing every second, network traffic classification becomes an essential step in intrusion detection.

The sophistication and complexity of contemporary cyber-attacks have begun to challenge the efficacy of the existing statistical and threshold approaches [4]. Machine Learning (ML) techniques have emerged as automatic and relevant solutions, particularly suited to intrusion detection and prevention within the realm of cloud security [5-7]. The efficacy of intrusion detection heavily relies on the accurate representation of features in network traffic. However, standard datasets available for such tasks are typically of high dimensionality and unbalanced, where each class label is not equally represented. This imbalance can potentially impact the performance of the classifier [8, 9].

When ML models are trained on such imbalanced datasets, the resultant predictive outcomes often favor dominant classes, leading to poor classification rates for minority classes and potentially misdirecting the predictive analysis [10]. Thus, it is imperative to develop intelligent systems capable of overcoming these biases when confronted with such data imbalances. Consequently, learning from imbalanced data has become a significant area of research over the past two decades. The concept of class imbalance has been extensively studied in diverse application areas, including but not limited to medical science [11], sentiment analysis [12], bioinformatics, intrusion detection, text mining [13], credit scoring, and fraud detection.

Resampling methods offer a potential solution to this problem by adjusting the class ratio to create a balanced dataset. When integrated with classification techniques, these resampling methods have the potential to substantially improve the intrusion detection and classification rates. Therefore, the exploration and optimization of such approaches forms the basis of this study.

This study endeavors to examine the impact of various resampling techniques on intrusion detection, including but not limited to Synthetic Minority Over-sampling Technique (SMOTE), Random Under Sampling (RUS), Random Over Sampling (ROS), and a combination of RUS-ROS and RUS-SMOTE. The initial approach involves the under-sampling of the majority class, followed by the application of the Synthetic Minority Oversampling Technique (SMOTE) for oversampling [2]. Ultimately, a combination of over-sampling and under-sampling techniques is employed to seek further enhancement in intrusion detection performance.

To assess the effectiveness of these resampling techniques, the Support Vector Machine classifier, renowned for its robust generalization capabilities and pattern recognition skills, is utilized [14]. The investigation employs the “CICIDS-2017” dataset, a benchmark dataset extensively detailed in the study of Vamsi Krishna et al. [15].

Evaluation metrics are subsequently applied to compare and contrast the performance of the classifier for each resampling technique. Although the intrusion detection rate is somewhat lower when compared to the combination of unbalancing techniques, it remains significantly higher than when the imbalance issue is left unaddressed. Consequently, the results obtained indicate an increase in performance.

The CICIDS2017 dataset, which exhibits a high-class imbalance, is composed of five days' worth of typical traffic and attacker traffic information sourced from the Canadian Institute of Cybersecurity. It encompasses benign traffic as well as the most recent common attacks, distributed among 14 classes, thereby closely mirroring real-world data. Previous research recommends the development of multi-class detector models on Tuesday, Wednesday, and Thursday mornings [13]. Therefore, the present study has chosen to focus on the dataset from Wednesday.

Given the aforementioned complexities inherent in high-dimensional, imbalanced datasets, this study aims to identify an effective resampling approach to balance the dataset, thereby enhancing classification accuracy. The contributions of this research are multifaceted and are outlined as follows:

(1) A novel framework, integrating hybrid pre-processing, is proposed to analyze and classify various benign network activities and malicious actions, leveraging the application of resampling techniques in the cloud domain.

(2) The highly unbalanced CICIDS2017 dataset is rebalanced using an array of resampling techniques, rendering it more meaningful and informative for model training.

(3) A suitable resampling technique for the Support Vector Machine (SVM) with a Radial Basis Function (RBF) classifier is suggested.

(4) The CICIDS2017 dataset is selected for performance metric evaluation and computational time analysis.

(5) The efficacy of hybrid pre-processing techniques is examined, with a focus on addressing the issue of disproportionate class distribution in the dataset.

(6) A review of the area of class imbalance is conducted, and to bolster the detection of rare attacks, Synthetic Minority Oversampling (SMOTE), Random Under Sampling, and Random Over Sampling techniques are employed.

(7) Efforts are made to balance the dataset, mitigating the negative effects of an imbalanced dataset on minority intrusion detection rates and other performance metrics, including recall rates.

When class distributions within a dataset are unequal, F1-Score and Recall metrics offer an optimal evaluation of the classification model. Higher metrics correlate with an improved ML model, yielding consistent grades. Performance measures underscore the importance of balanced datasets in optimizing intrusion detection systems, highlighting the performance degradation caused by imbalanced datasets.

The remainder of the paper is structured as follows: Section 2 encompasses related studies conducted by several researchers in the field. Section 3 presents the proposed method. Section 4 is dedicated to results and discussion, and finally, Section 5 draws conclusions from the experimental findings.

2. LITERATURE REVIEW

The issue of class imbalance necessitates the utilization of resampling strategies. The confluence of resampling techniques with supervised classifiers becomes critical in the realm of network intrusion detection. Comprehensive surveys and analyses of diverse studies indicate that imbalanced learning contributes significantly to the performance of intrusion detection methodologies. This section provides a succinct overview of select previous studies. Recent developments in the field of resampling methods are also discussed.

A novel hybrid framework, "ImmuneNet", is described in the study of Kumaar et al. [16] that aims at fortifying the security of patient records in healthcare systems. This

framework is a synthesis of deep learning and feature engineering techniques. It employs an array of oversampling methods and hyper-parameter optimization strategies to enhance accuracy and performance. The "ImmuneNet" framework was subjected to rigorous testing against various benchmark datasets, including CIC-IDS-2017, CI-IDS-2018, and Bell DNS 2021. Additionally, four different machine learning algorithms were evaluated. The authors concluded that "ImmuneNet" demonstrated superior performance, achieving an accuracy of 99.19% and an ROC-AUC of 99.2%.

In a study by Ustebay et al. [17], a fusion of recursive feature elimination via Random Forest and Deep Multilayer Perceptron (DMLP) is proposed for Intrusion Detection Systems (IDS). This research aims to tackle the challenges posed by big data, identifying the impact of diverse attributes from the dataset and determining the most informative features that meaningfully represent the data. In the Random Forest (RF) methodology, various tree structures are employed to discard non-essential features from a total of 80. Based on the graphical representations, 10 key features are selected. The truncated data is then subjected to binary classification using DMLP, resulting in an accuracy of 89%.

In another research effort [18], an ensemble learning approach is proposed, constituting a resilient and efficacious intrusion detection framework using eXtreme Gradient Boosting (XGBoost) alongside an embedded feature selection methodology. Ensemble-based Gradient Boost Trees are utilized as a filter method, evaluating each feature based on its significance. The embedded feature selection approach unfolds in three phases, examining a subset of features by employing filter methods, wrapper methods, or hybrid methods. Subsequently, the imbalance issue is addressed by amalgamating similar types of attacks into a single category, resulting in categories of Normal, WebAttack, Infiltration, BruteForce, Dos, Botnet, PortScan, and DDoS types. This methodology is tested using the CICIDS 2017 dataset and is evaluated for both binary and multi-class classification problems. The XGBoost classifier is finally applied as an evaluator, achieving an accuracy of 99.86% and 99.90% for binary and multi-class classification, respectively.

In the face of burgeoning data volumes and expansive Internet connectivity, the significance of intrusion detection in safeguarding our infrastructure and national security cannot be overstated. A model employing machine learning techniques for network intrusion detection is proposed by Tauscher et al. [19]. This model is bifurcated into two stages: the initial stage is dedicated to distinguishing normal from suspicious behaviors, while the subsequent stage classifies specific attacks. To address the predicament of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is employed by the authors. The model's performance is evaluated with eight classifiers, among which an unsupervised autoencoder-based model rendered the most effective performance. All experiments were conducted using the NSL-KDD dataset. However, a significant limitation of this study is the exclusive reliance on a single technique to balance the dataset.

In the study of Fu et al. [2], a network intrusion detection system is introduced that addresses challenges such as data imbalance and low detection rate accuracy in cloud computing. The ADASYN oversampling algorithm is utilized to increase the number of samples in minority class labels, thereby addressing data imbalance. Furthermore, the model's generalization capability is enhanced, and the network structure is refined by integrating the channel attention mechanism with bidirectional LSTM networks. Comparison with other models in the literature suggests that the proposed DLNID yields superior classification results. The model is evaluated using the publicly accessible benchmark dataset NSL-KDD, outperforming the other methods compared, with an accuracy of 90.73% and an F1-score of 89.65%.

To fortify defenses against emerging threats, an advanced anomaly intrusion detection system is proposed by Elmasri et al. [20]. The proposed model leverages both KNN and LOF algorithms, supplemented by an enhanced version utilizing PCA. In terms of detecting novel attacks, both models demonstrate considerable proficiency, achieving detection rates of 88.3% and 90.54%, respectively. Moreover, a significant reduction in time complexity is observed across all models.

In another comprehensive study by Razan Abdulhammed et al. [21], two feature dimensionality reduction methods, Auto-encoder (AE) and Principal Component Analysis (PCA), were employed. The authors tested various classifiers, including Random Forest (RF), Bayesian Network, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA), using a subset of the CICIDS 2017 dataset. AE was used to reduce the dimensionality of the CICIDS 2017 dataset from 81 to 59, while PCA was used to further reduce it to 10. In both instances, Random Forest exhibited superior accuracy compared to the other classifiers.

A framework for Network Intrusion Detection Systems (NIDS) designed to combat a wide range of contemporary and emerging threats is proposed by Ahmed et al. [22]. Five machine learning algorithms, namely Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Artificial Neural Networks (ANN), were evaluated for their effectiveness in classifying attacks. The UNSW-NB15 dataset was used to assess the performance of the proposed framework. Feature selection and SMOTE resampling methods were employed in conjunction with these algorithms to balance the classification labels. Accuracies were reported for classification models both with and without feature selection, following the handling of class imbalance. In both scenarios, the Random Forest (RF) classifier demonstrated superior performance, achieving 95.1% accuracy when used in conjunction with SMOTE and PCA. However, time complexity was not considered in this study.

In the context of IoT networks, where security and privacy are paramount, a multi-stage classification system is proposed by Qaddoura et al. [23]. This system comprises three stages: dataset size reduction using k-means clustering, dataset balancing using the SMOTE oversampling technique, and the subsequent classification of the balanced data. A comparative analysis of various classifiers in the final stage revealed SVM-SMOTE to be the most effective.

The imbalanced dataset CIDDS-001 was addressed using various techniques in the study of Abdulhammed et al. [24]. The Synthetic Minority Reconstruction Technique (SMRT) in combination with a Variational Autoencoder (VAE) was applied to the data for classification. A multitude of sampling methods were used to address class imbalance, including the down-sampling of majority classes and up-sampling of minority classes. Following this, the dataset underwent analysis using various machine/deep learning algorithms, such as random forest and deep neural networks, to evaluate progress in attack detection. However, computational overhead was not considered in this study.

Other studies have employed deep learning with different tuning hyperparameters to mitigate imbalance issues and enhance classification accuracy on the CICIDS (2017) dataset, achieving 97.7% accuracy [25]. An alternative approach proposed by Yang et al. [26] uses the Self-Paced Ensemble and Auxiliary Classifier Generative Adversarial Networks (SPE-ACGAN) to address the imbalance issue of sample classes, leading to an increase in precision (82.23%), recall (82.54%), and F1-score (82.38%) on the CICIDS dataset.

In the majority of the aforementioned studies, emphasis is placed on the utilization of an oversampling technique, specifically Synthetic Minority Oversampling Technique (SMOTE), to address the issue of class imbalance. However, there exists a potential drawback to oversampling techniques, particularly in multi-class classification scenarios, where they may exacerbate the underrepresentation of minority classes. From the extant literature, only a handful of studies have embarked on a comparative analysis of various resampling techniques, with a predominant focus on datasets related to networks and the Internet of Things (IoT).

The current study builds upon the foundation laid by preceding researchers, proposing a hybrid preprocessing model designed to reduce the number of majority samples and augment the number of minority samples. This approach is aimed at mitigating the imbalance in the training set, thereby enhancing the classification accuracy and performance metrics of the intrusion detection system in a cloud environment. In addition, it targets the reduction of potential underrepresentation of minority classes.

To achieve this objective, three balancing techniques are considered: Synthetic Minority Oversampling (SMOTE), Random Under Sampling (RUS), and Random Over Sampling (ROS). Various combinations of these techniques are tested in an effort to achieve balanced classes. The proposed model is evaluated using the publicly available benchmark dataset CICIDS-2017, and performance metrics are assessed with the SVM-RBF classifier.

This endeavor extends the body of knowledge on the subject, emphasizing the importance of class balance in the performance of intrusion detection systems, and exploring novel approaches to address this crucial aspect.

3. MEHTODOLOGY

In this segment, a new model is introduced for intrusion detection to know and analyze the effectiveness of two resampling methods, two hybrid models (combination of these resampling methods) and also without resampling. In order to evaluate these models from the perspective of performance, the SVM-RBF classifier is identified and the model is implemented using a benchmark data set that comprises of all attacks (CICIDS-2017). This methodology is divided into three phases: A) Data preprocessing, B) Combating the Imbalance and C) Classification. The following flow chart in Figure 1 depicts the proposed framework.

3.1 Data preprocessing

This particular task involves pooling and segregating the dataset and scrapping all the features which are not needed for analysis. The original dataset contains 80 features. Present study adopts Several prominent preprocessing techniques form [27] including “data cleaning, transformation, and normalization techniques”, on the CICIDS2017 Wednesday dataset generated by Sharafaldin et al. [28]. As a result, the following features have been droped, namely “Bwd PSH Flags, Fwd URG Flags, Bwd URG Flags, CWE Flag Count, Fwd Avg Bytes, Fwd Avg Packets, Fwd Avg Bulk Rate, Bwd Avg Bytes, Bwd Avg Packets, Bwd Avg Bulk Rate, Bwd Avg Bytes, Bwd Avg Packets”. By the end of the preprocessing phase, the resulting dataset contains 68 features and the corresponding class distribution is shown in Table 1. After which, the benchmark data set is segmented into a training data set that consists of 80% and a test data set that consists of 20%.

Table 1. Class wise distribution of CICIDS2017 dataset after data preprocessing

Category	Class	Number of Records
Anomaly	DoS GoldenEye	10,293
	DoS Hulk	230,124
	DoS Slowhttptest	5,499
	DoS slowloris	5,796
	Heartbleed	11
Total Anomaly data	---	2,51,723
Normal	Benign	4,39,972
Total	---	6,91,695

3.2 Combating the imbalance

In real-world datasets, unbalanced datasets are common. Several sampling techniques can be used to eliminate samples in the overlapping region to improve classification performance and deal with unbalanced class distributions, such as over-sampling or under-sampling. Recently, Machine learning has become increasingly popular with the use of hybrid methods. In this study, the combination of these techniques is used to improve classification performance. The following models are tested.

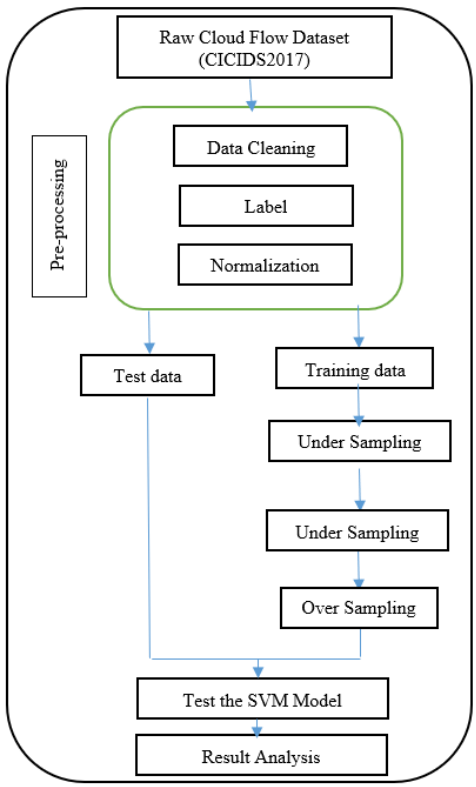

Figure 1. Flowchart of the proposed methodology

- No resampling (NR).
- Synthetic Minority Over-sampling Technique (SMOTE).
- Random Under Sampling (RUS).
- Random Under Sampling and Random Over Sampling (RUS+ROS).
- Random Under Sampling and SMOTE (RUS+SMOTE).

Random under-sampling is a process of reducing the number of samples randomly in the majority classes which automatically leads to reducing the training time. But there is the possibility of losing valuable data.

SMOTE is an oversampling technique which generates duplicate samples within the feature space of the training dataset that is relatively close to the minority class before training the model.

This method computes the Euclidian distances between the feature vectors of each minority class sample using the nearest neighbors which are located. As a result, synthetic samples are generated between each sample and its neighbors.

Consequently, it may be possible to introduce false information to the model, which may lead to overfitting. This increases the amount of time spent on training as well. The advantage of oversampling is that we will not lose any valuable information since any of the data points are not deleted.

A hybridization technique that involves both under-sampling and over-sampling techniques can be beneficial for improving the performance of the classifier by limiting the count of overlapping samples in the feature space and reducing the training period. It is recommended by Das et al. [29] to apply first under-sampling before oversampling to reduce the effects of overlapping classes. Consequently, the model's generalization ability can be further enhanced and the disproportion in the network data can be solved to some considerable amount. For this reason, the two hybrid models that combine RUS and ROS(RUS+ROS) and RUS and SMOTE(RUS+SMOTE) are tested to study the effectiveness of these combinations.

In all of the above models to balance the data points and to generalize the ratio, the sampling strategy is taken as an average of the total count of samples versus the count of classes in the training dataset is 80698. The C and Gamma parameters of SVM-RBF are set with default values of scikit library and all the experiments were conducted using python programming in Jupyter notebook.

The Table 2 and Figure 2 depict samples taken before and after resampling with SMOTE, RUS, RUS+ROS, and RUS+SMOTE. According to Table 2, Heartbleed was found to have 8 instances, while DoS Slowhttptest had 3823 before resampling. Oversampling would make a big difference in these two attacks. SMOTE has the highest number of samples, based on the observation of the sample count. Because of this, the SMOTE+SVM-RBF classifier requires a longer training time than other classifiers. With RUS, the number of benign samples is reduced to 84%, while DoS Hulk samples are reduced to 50%. However, other classes had a lower number of samples as collated into these two classes. The consequences are not balanced. RUS+ROS equalized the samples in each class. With RUS+SMOTE, the samples for DoS Golden Eye, DoS Slowhttptest, DoS Slowloris, and Heartbleed were made equal.

3.3 Classification using SVM-RBF

SVM classifier is used in this module with RBF kernel to classify multi-category data. The SVM-RBF is more appropriate when compared to other existing classifier for multiclass classification, according to the existing literature. The output of the combating the class imbalance phase is input to this classifier. Multi-class classifications in SVM can be handled using two techniques: "one-versus-one" (OVO) and "one-versus-all" (OVA). As per the study of Khan [4], fifteen binary models are needed for the OVO or pairwise classification of the CICIDS2017 dataset with six classes. Moreover, it requires more computation than the OVA approach, which requires only six models to distinguish all classes [4, 5]. It is for these reasons that the OVA approach is considered when conducting experiments. Further, the standard metrics which helps to quantify the performance of classifier to be particular are accuracy, precision, recall, F1-Score and time complexity are considered and analyzed. In this experiment, the Intel Core i5 processor with 1.80 GHz and 8GB RAM was used, along with Windows 10 as the operating system.

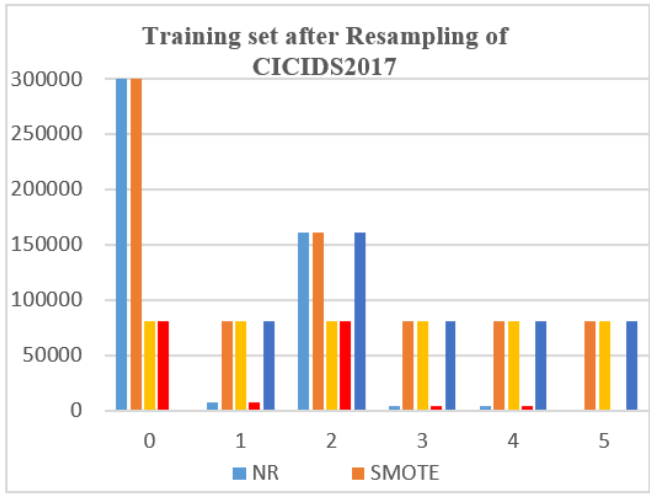

Figure 2. Training set after Resampling of CICIDS2017

Table 2. Training dataset class distribution after resampling of CICIDS2017

Category	Label	NR	SMOTE	RUS	RUS+ROS	RUS+SMOTE
Benign	0	307,927	307,927	80,698	80,698	8
DoS Golden Eye	1	7,211	80,697	7,211	80,698	80,697
DoS Hulk	2	161,096	161,096	80,698	80,698	161,096
DoS Slowhttptest	3	3,823	80,697	3823	80,698	80,697
DoS slowloris	4	4,121	80,697	4121	80,698	80,697
Heartbleed	5	8	80,697	8	80,698	80,697

4. RESULTS AND DISCUSSION

The effect of attack detection rate discussed in the previous section have been assessed and analyzed. The comparison of various classification metrics obtained from the SVM-RBF classifier is done. The class distribution in the CICIDS2017 dataset is unbalanced. Because of this, accuracy alone is not the appropriate metric to appraise best learning algorithms. The accuracy may be great if the majority class is classified precisely, even if the rare classes are incorrectly classified. A better option to compare sampling techniques' performance is to examine the classifier's precision and recall along with accuracy.

The following observations are from Table 3 and Figure 3 regarding accuracy. It has been found that the RUS+SMOTE achieves best in attack classification with an accuracy of 99.63%. Then the next order follows by SMOTE and RUS and exhibit more or less similar behaviors for both training and testing.

The result of Precision, Recall and F1-score for various models illustrated in Table 3 and Figure 4, the Precision of the RUS+SMOTE shows highest performance with 0.99. With 0.92, SMOTE is in second, followed by RUS, RUS+ROS, both of which are equal.

Table 3. Performance comparison of various sampling methods

Model	Training- Accuracy	Testing -Accuracy	Precision	Recall	F1-Score	Training -Time (s)	Testing –Time(s)
NR	96.15	96.2	0.79	0.68	0.72	24,690.51	1,123.72
SMOTE	97.19	97.2	0.92	0.98	0.95	19,960.59	893
RUS	96.93	96.93	0.8	0.98	0.86	**868.688**	300.85
RUS +ROS	97.91	95.22	0.8	0.98	0.86	4,574.33	594.7
RUS+SMOTE	**99.37**	**99.63**	**0.99**	0.94	**0.96**	2215	**112.59**

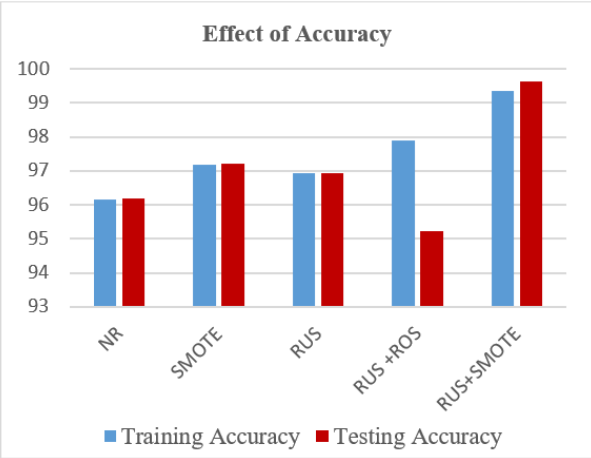

Figure 3. Accuracy fir the training and testing for various resampling methods

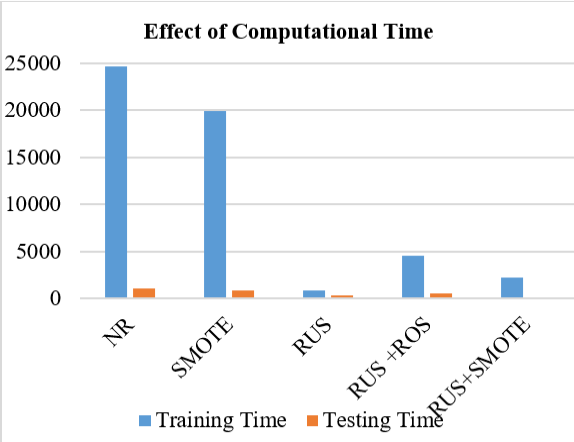

Figure 4. Effect of computational time for different resampling models

The Recall values of SMOTE, RUS and RUS+ROS are more elevated in comparison with various methods with the result 0.98. Next in line is RUS+SMOTE with a marginal difference of 0.04(0.94). But in case of F1-score the RUS+SMOTE performed best with a value of 0.96 in comparison to other models. The next utmost value 0.95 is by SMOTE escorted by RUS, RUS+ROS with a value 0.86. Despite an increase of 0.98% into accuracy of RUS+ROS over RUS, all the three results of precision, recall and F1-score are equal indicates that results are biased.

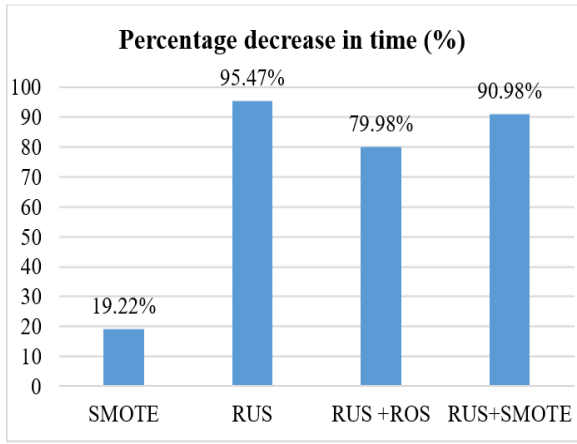

Figure 5. Percentage decrease in computational time for various resampling models

Table 4. Percentage gain of various resampling models

Model	Percentage Decrease in Time (%)
SMOTE	19.22
RUS	95.47
RUS+ROS	79.98
RUS+SMOTE	90.98

From Table 4, Figure 4 and Figure 5, it is clear that the RUS algorithm presents the lowest average computational time but the intrusion detection capability has not been improved in terms of false positive and minority classes. Although SMOTE improves the performance of the base classifier SVM-RBF, it needs a lot of training time. It generates synthetic samples for minority classes and accounts for the majority of the computational time.

Overall, it is evident from Table 3 that the balanced dataset has improved the performance of intrusion detection by detecting positive instance as itself, indicating very few false

positives in the results. The class-wise performance can be explained using F1-score that yields to 96% whereas the overall performance is represented in terms of accuracy (99.63%) is also the good indicator for balanced dataset. Thus, it is clear from the results that the RUS+SMOTE performs the best in intrusion detection compared to the studies of Khandekar et al. [30, 31]. Although our results are still lower than those in the study of Mbow et al. [32] in terms of intrusion detection rate and F1-score.

However, this study was focused on only SVM-RBF kernel and did not consider the other kernels. Additionally, analysis of feature selection can be done to select the subset of original dataset before and after resampling methods to mitigate the false positive rate.

5. CONCLUSION

This experimental analysis aims to examine, various resampling methods, to mitigate the class imbalance problem in NIDS using SVM-RBF classifier. The SVM-RBF classifier receives input data from different sampling methods adopted after generating the synthetic data to balance the class distribution.

From the results obtained, it is observed that the combination of SMOTE+RUS yields high intrusion detection results (99.63%) compared to other models including without applying resampling i.e., 96.15%. But in case of Recall and Computational time the RUS method performs well, with marginal difference in Recall value (0.04 %) and gain in computational time is 4.49%. Based on these results to suggest that the combination of two methods, over sampling method SMOTE and under sampling method RUS are most effect of the data balancing on classification performance, even though it takes marginal amount of training time because the SMOTE add additional records to minority classes. This process yields better accuracy with high detection rate of minority classes. Future research will focus on studying the comparative analysis of various kernels of SVM classifier. In addition, the effect of feature selection to select the subset of original dataset before and after resampling methods can be tested by the proposed approach.

REFERENCES

[1] Ogwara, N.O., Petrova, K., Yang, M.L. (2022). Towards the development of a cloud computing intrusion detection framework using an ensemble hybrid feature selection approach. Journal of Computer Networks and Communications, 2022: 1-16. https://doi.org/10.1155/2022/5988567

[2] Fu, Y., Du, Y., Cao, Z., Li, Q., Xiang, W. (2022). A Deep Learning Model for Network Intrusion Detection with Imbalanced Data. Electronics, 11(6): 898. https://doi.org/10.3390/electronics11060898

[3] Narisetty, N., Kancherla, G. R., Bobba, B., Swathi, K. (2021). Hybrid Intrusion detection method based on constraints optimized SAE and grid search based SVM-RBF on cloud. International Journal of Computer Networks and Applications, 8(6): 776-787. https://doi.org/10.22247/ijcna/2021/210725

[4] Khan, M.A. (2021). HCRNNIDS: hybrid convolutional recurrent neural network-based network intrusion detection system. Processes, 9(5): 834. https://doi.org/10.3390/pr9050834

[5] Tulasi Bhavani, T., Rao, M.K., Reddy, A.M. (2020). Network intrusion detection system using random forest and decision tree machine learning techniques. Advances in Intelligent Systems and Computing, Rajasthan, India, 637-643. https://doi.org/10.1007/978-981-15-0029-9_50

[6] Krishna Anne, V.P., Rajasekhara Rao, K. (2017). Standards and analysis of intrusion detection-based system: A comparative study. Ponte, 73(2): 87-97. https://doi.org/10.21506/j.ponte.2017.2.7

[7] Jadhav, A.D., Pellakuri, V. (2019). Performance analysis of machine learning techniques for intrusion detection system. In 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, pp. 1-9. https://doi.org/10.1109/ICCUBEA47591.2019.9128917

[8] Zhang, Y.Z. (2018). Deep generative model for multi-class imbalanced learning. Open Access Master's Theses. https://doi.org/10.23860/thesis-zhang-yazhou-2018

[9] Chawla, N.V. (2009). Data mining for imbalanced datasets: An overview. In Data Mining and Knowledge Discovery Handbook, New York, NY, pp. 875-886. https://doi.org/10.1007/978-0-387-09823-4_45

[10] Bagui, S., Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. Journal of Big Data, 8(1): 1-41. https://doi.org/10.1186/s40537-020-00390-x

[11] Krawczyk, B., Galar, M., Jeleń, Ł., Herrera, F. (2016). Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Applied Soft Computing, 38: 714-726. https://doi.org/10.1016/j.asoc.2015.08.060

[12] Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., Wang, X. (2015). Word embedding composition for data imbalances in sentiment and emotion classification. Cognitive Computation, 7(2): 226-240. https://doi.org/10.1007/s12559-015-9319-y

[13] Munkhdalai, T., Namsrai, O.E., Ryu, K.H. (2015). Self-training in significance space of support vectors for imbalanced biomedical event data. BMC Bioinformatics, 16(7): 1-8. https://doi.org/10.1186/1471-2105-16-S7-S6

[14] Narisetty, N., Kancherla, G.R., Bobba, B., Swathi, K. (2021). Performance analysis of different activation and loss functions of stacked autoencoder for dimension reduction for NIDS on cloud environment. International Journal of Engineering Trends and Technology, 69(4): 169-176. https://doi.org/10.14445/22315381/IJETT-V69I4P224

[15] Vamsi Krishna, K., Swathi, K., Rama Koteswara Rao, P., Basaveswara Rao, B. (2022). A detailed analysis of the CIDDS-001 and CICIDS-2017 datasets. In Pervasive Computing and Social Networking, Salem, India, pp. 619-638. https://doi.org/10.1007/978-981-16-5640-8_47

[16] Kumaar, M.A., Samiayya, D., Vincent, P.D.R., Srinivasan, K., Chang, C.Y., Ganesh, H. (2021). A Hybrid framework for intrusion detection in healthcare systems using deep learning. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2021.824898

[17] Ustebay, S., Turgut, Z., Aydin, M.A. (2018). Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier. In 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), Ankara, Turkey, pp. 71-76.

https://doi.org/10.1109/IBIGDELFT.2018.8625318
[18] Mokbal, F., Dan, W., Osman, M., Ping, Y., Alsamhi, S. (2022). An efficient intrusion detection framework based on embedding feature selection and ensemble learning technique. The International Arab Journal of Information Technology, 19(2): 237-248. https://doi.org/10.34028/iajit/19/2/11
[19] Tauscher, Z., Jiang, Y., Zhang, K., Wang, J., Song, H. (2021). Learning to detect: A data-driven approach for network intrusion detection. In 2021 IEEE International Performance, Computing, and Communications Conference (IPCCC), Austin, TX, USA, pp. 1-6. https://doi.org/10.1109/IPCCC51483.2021.9679415
[20] Elmasri, T., Samir, N., Mashaly, M., Atef, Y. (2020). Evaluation of CICIDS2017 with qualitative comparison of Machine Learning algorithm. In 2020 IEEE Cloud Summit, Harrisburg, PA, USA, pp. 46-51. https://doi.org/10.1109/IEEECloudSummit48914.2020.00013
[21] Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., Abuzneid, A. (2019). Features dimensionality reduction approaches for machine learning based network intrusion detection. Electronics, 8(3): 322. https://doi.org/10.3390/electronics8030322
[22] Ahmed, H.A., Hameed, A., Bawany, N.Z. (2022). Network intrusion detection using oversampling technique and machine learning algorithms. PeerJ Computer Science, 8: e820. https://doi.org/10.7717/peerj-cs.820
[23] Qaddoura, R., Al-Zoubi, A. M., Almomani, I., Faris, H. (2021). A multi-stage classification approach for iot intrusion detection based on clustering with oversampling. Applied Sciences, 11(7): 302. https://doi.org/10.3390/app11073022
[24] Abdulhammed, R., Faezipour, M., Abuzneid, A., AbuMallouh, A. (2018). Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic. IEEE Sensors Letters, 3(1): 1-4. https://doi.org/10.1109/LSENS.2018.2879990
[25] Choraś, M., Pawlicki, M. (2021). Intrusion detection approach based on optimised artificial neural network. Neurocomputing, 452: 705-715. https://doi.org/10.1016/j.neucom.2020.07.138
[26] Yang, H., Xu, J., Xiao, Y., Hu, L. (2023). SPE-ACGAN: A resampling approach for class imbalance problem in network intrusion detection systems. Electronics, 12(15): 3323. https://doi.org/10.3390/electronics12153323
[27] Narisetty, N., Kancherla, G. R., Bobba, B., Swathi, K. (2021). Investigative study of the effect of various activation functions with stacked autoencoder for dimension reduction of NIDS using SVM. International Journal of Advanced Computer Science and Applications, 12(5): 152-161. http://doi.org/10.14569/IJACSA.2021.0120519
[28] Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. International Conference on Information Systems Security and Privacy, 1: 108-116. https://doi.org/10.5220/0006639801080116
[29] Das, B., Krishnan, N.C., Cook, D.J. (2014). Handling imbalanced and overlapping classes in smart environments prompting dataset. Data mining for service, 199-219. https://doi.org/10.1007/978-3-642-45252-9_12
[30] Khandekar, V.S., Shrinath, P. (2022). Ensemble Model for Multiclass Imbalanced Data Using Cluster Computing of Spark. https://doi.org/10.21203/rs.3.rs-1981706/v1
[31] Govindarajan, M. (2022). Effective intrusion detection system using classifier ensembles. Ingénierie des Systèmes d'Information, 27(1): 151-156. https://doi.org/10.18280/isi.270118
[32] Mbow, M., Koide, H., Sakurai, K. (2022). Handling class Imbalance problem in Intrusion Detection System based on deep learning. International Journal of Networking and Computing, 12(2): 467-492. https://doi.org/10.15803/ijnc.12.2_467