IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 November 2023, accepted 4 December 2023, date of publication 12 December 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3341755

RESEARCH ARTICLE

Learning From Few Cyber-Attacks: Addressing the Class Imbalance Problem in Machine Learning-Based Intrusion Detection in Software-Defined Networking

SEYED MOHAMMAD HADI MIRSADEGHI[1], HAYRETDIN BAHSI[1,2], RISTO VAARANDI[1], AND WISSEM INOUBLI[3]

[1]Centre for Digital Forensics and Cyber Security, Department of Software Science, Tallinn University of Technology, 19086 Tallinn, Estonia
[2]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, USA
[3]CNRS, UMR 8188, Centre de Recherche en Informatique de Lens (CRIL), Artois University, 62300 Lens, France

Corresponding author: Seyed Mohammad Hadi Mirsadeghi (seyed.mirsadeghi@taltech.ee)

This work was supported by the EU Horizon2020 project MariCybERA under Agreement 952360.

ABSTRACT The class imbalance problem negatively impacts learning algorithms' performance in minority classes which may constitute more severe attacks than the majority ones. This study investigates the benefits of balancing strategies and imbalanced learning approaches on intrusion data from Software Defined Networking (SDN). Although the research community has covered the imbalance problem in machine learning-based intrusion detection, addressing this problem in SDN is novel and powerful. Addressing the class imbalance problem over InSDN (the only publicly available SDN intrusion detection dataset as of recent) is of significant impact on future research in the area of intrusion detection in SDN. We address the class imbalance problem through data-level and classifier-level techniques. Our research objective is to determine suitable methods of addressing the class imbalance problem in machine learning-based intrusion detection in SDN. We propose custom deep learning architectures based on GANs and Siamese Neural Networks for generative modeling and similarity-based intrusion detection. This paper provides benchmarking results from classification with Random Oversampling (ROS), SMOTE, GANs, weighted Random Forest, and Siamese-based one-shot learning. We have found that Random Forest (RF) outperforms deep learning models in the classification of minority class instances. This supports the notion that RF can handle class imbalance well. We also observe that widely-used balancing techniques, ROS and SMOTE, drastically decrease the False Positive Rate (FPR) but increase the False Negative Rate (FNR) in the classification of minority classes. Conclusively, while data-level methods improve classification performance over deep learning models, they, in fact, degrade RF's performance, i.e. cause higher numbers of false predictions. Therefore, RF does not need additional balancing strategies to get higher performance. Although this work addresses the class imbalance problem in SDN intrusion data, it provides a well-designed benchmark that can be exemplary for any network intrusion detection data. Thus, it may have a significant impact on future studies in this respective domain.

INDEX TERMS Class imbalance problem, machine learning, deep learning, cyber intrusion detection, software-defined networking.

I. INTRODUCTION

Intrusion detection is a critical security function for identifying cyber attacks in real time and initiating the appropriate responses to threats. Common solutions depend on attack signatures resulting in a limitation in detecting future variations or enhancements of existing attack types. Machine learning-based approaches promise to detect new attack types by learning from statistical patterns in network traffic data.

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

Most intrusion datasets incorporate severe imbalance ratios between classes. InSDN [12] (a novel SDN intrusion dataset) includes more cases of DoS, DDoS, and Probe attacks than all other classes combined. Severe class imbalance usually leads to poor performance of learning algorithms in generalization of classes in minority. For example, a recent study [32] shows that Decision Trees handle the class imbalance problem better than neural networks in the AWID WiFi Intrusion dataset. Another study shows that Deep Learning models, in particular, fail to learn the generalizable features of minority classes and lean more towards the majority class [25]. In learning over imbalanced InSDN dataset [38] machine learning-based IDS yields adequate performance in detecting majority classes such as Normal, DDoS, DoS, and Probe, which can be easily detected by the contemporary detection technologies. However, such a detection system creates high numbers of false negatives and false positives in minority classes, i.e. exploitation attack types such as U2R (privilege escalation), Web-attack, and Botnet, which is all the more important to detect as they usually appear in the form of very few packets over the network and may have more severe consequences, e.g., acquiring control of target system, data exfiltration, et c. In some cases, attackers seek to masquerade exfiltration traffic by mimicking innocuous protocols such as DNS, a common choice of amplifier in DDoS attacks [1], [6], [35].

Machine Learning algorithms offer value in the detection of Cyberattacks in software-defined settings. Minority attack classes, i.e. classes that are underrepresented, are usually challenging for a learning algorithm to detect. for example, exfiltration traffic rates may be hidden in high-volume flow traces of DDoS attack. The task of imbalanced learning can be viewed as alleviating a learning algorithm's bias towards majority class instances. Learning algorithms generally build classification models based on maximum accuracy, which may lead to biased classification towards the majority class and misclassification of the minority class instances [11], [31], [54]. Poor classification performance over imbalanced datasets is caused not always solely by class imbalance but also by class overlap. For example, a linearly separable dataset can be perfectly classified by a typical classification algorithm regardless of how skewed the class distribution is [3]. Class overlap occurs when multiple classes share the same region in the data space as seen by the classifier. It is in the presence of class overlap that even a balanced dataset can be challenging for a learning algorithm. In fact, class overlap shows the highest negative impact among potential factors including class imbalance [11], [53]. Classification becomes more difficult with the class overlap problem when class imbalance is also present in the data, and vice versa [17].

Learning systems trained on imbalanced datasets typically exhibit bias towards majority groups. Cost-sensitive learning [59], for example, assigns a varying penalty to each class and seeks to minimize the misclassification error. Weighted Decision Trees are another example of an effective approach to learning from data distributions where minority class instances are given higher priority. Oversampling of minority classes prior to training is an effective method of learning from imbalanced data distributions. Synthetic Minority Oversampling TEchnique (SMOTE) [8] is an over-sampling approach in which the minority class is over-sampled by creating synthetic examples along the line segments joining any/all of the k minority class nearest neighbors.

Even though the class imbalance problem was first recognized three decades ago, it remains a challenge in an evolving research area. The class imbalance problem has been addressed mainly in two ways [21]: by (1) employing dataset-balancing techniques prior to classification (data-level method) and (2) incorporating imbalance-aware classification methods (classification-level method). The former alters the sizes of training datasets by adding or reducing some instances in order to achieve balanced learning. The latter proposes models that learn directly from an unbalanced dataset without modifying it.

Even though the class imbalance problem has been discussed over well-known intrusion detection datasets such as NSL-KDD [50] and UNSW-NB15 [42], this problem has not been addressed over publicly available SDN intrusion data. This is perhaps because InSDN [12] is the first publicly available intrusion dataset that was generated in an SDN testbed. Like most intrusion detection datasets, InSDN suffers from the class imbalance problem.

This study aims to identify and compare suitable methods to address the class imbalance problem in SDN intrusion data as well as provide comprehensive benchmarking that compares various data-balancing and classifier-level methods for a multi-class classification problem in detecting intrusions in SDNs. We ask the question: how effective are shallow and deep learning-based data-level and classifier-level methods in addressing the class imbalance problem in machine learning-based intrusion detection in SDN? More specifically, while evaluating data-balancing strategies, we selected widely-used naive approaches such as Random Oversampling (ROS) and Random Undersampling (RUS) in addition to more complex approaches such as SMOTE and Generative Adversarial Networks (GANs). As data balancing should be complemented by a learning model, we investigated the performance of two deep-learning models with varying parameter sizes (i.e., Multi-Layer Perceptron (MLP) models with 6 layers and 10 layers) and a shallow model (i.e., Random Forest) with the above-mentioned balancing strategies. We selected two different classifier-level methods: one-shot learning and weighted random forest. The overall detection performance of the model is evaluated by the multi-class classification metrics Macro-F1 and Micro-F1. Macro-F1 better reflects the model performance as it equally considers the minority classes. Measures of accuracy, recall, precision, and F1 scores are utilized for assessing the model performance on each attack type.

This work presents several contributions:

- We provide a comprehensive benchmarking study that compares data-level and classifier-level strategies for the SDN dataset.

- We present detailed experimental results that compare how data-level methods perform with deep learning and shallow learning models
- We explain the impact of each model choice on majority and minority classes in a detailed way

To our knowledge, a detailed benchmarking study that assesses data balancing strategies within SDN intrusion data has not been conducted before. The research community can also benefit from our benchmarking design while evaluating network intrusion detection datasets in similar domains.

The remainder of the paper is organized as follows: Section II describes some key findings in the literature. Section III describes the materials and methods employed in this work and Section IV presents the results. Section V discusses the main findings and limitations of this work, and Section VI concludes the paper.

II. RELATED WORK

The class imbalance problem has been widely discussed in the literature. A comprehensive review of research in learning from imbalanced data [21] discusses ROS and RUS, synthetic data generation, cost-sensitive learning, and active learning as effective methods to deal with the class imbalance problem. It has been demonstrated that the effect of class imbalance is detrimental to classification performance over three benchmark datasets, i.e. MNIST, CIFAR-10, and ImageNet [5]. A comprehensive review of the development of research in learning from imbalanced data [21] provides a critical review of the nature of the problem, state-of-the-art solutions, and assessment metrics. Methods of dealing with the class imbalance problem are studied for classical machine learning models [7], [24], [36].

A study on the effect of the class imbalance problem on ml-based network intrusion detection [52] concludes that downsampling coupled with upsampling and SMOTE is the best re-sampling technique over the NSL-KDD dataset. The authors propose an ensemble model that achieves the highest performance. Downsampling, upsampling, and SMOTE are simple yet effective methods. The strength of sampling methods is their simplicity. While SMOTE produces data points that belong to the original distribution, upsampling and downsampling may not provide expected results sometimes even degrading learning models' performance due to duplication and underrepresentation, respectively.

This study [51] uses SMOTE to develop a rich training set prior to classification via random forest. They conclude that this approach reduced the time required to build the model and increased the detection rate for minority classes substantially.

A common solution to the class imbalance problem is the oversampling of minority classes. A combined Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO) technique has been proposed [16] for the two-class imbalanced classification problem that utilizes the radial basis function (RBF) classifier. Data augmentation using Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) [41] has been proposed as an effective approach to dealing with the class imbalance problem in deep learning. While deep learning-based sampling methods based on Auto Encoders and Generative Adversarial Networks have proven effective in producing realistic synthetic data, the complexity of these methods is their weakness. These models require large training sets and more optimization time compared to simpler methods such as oversampling and SMOTE.

This paper [2] focuses on resampling methods as the solution to the class imbalance problem over benchmark network intrusion datasets such as KDD99, UNSW-NB15, UNSW-NB17, and UNSW-NB18. They conclude that oversampling and undersampling both increase the measure of recall significantly in the case of severely imbalanced data. A novel Difficult Set Sampling Technique (DSSTE) algorithm [30] has been proposed to tackle the class imbalance problem that utilizes the Edited Nearest Neighbor (ENN) algorithm to divide the imbalanced training set into separate buckets and then uses the k-means algorithm to compress the majority class samples. They combine majority and minority class instances in the difficult set in order to perform data augmentation. Experimental results on classic intrusion datasets NSL-KDD and CSE-CIC-IDS2018 show that their algorithm outperforms methods such as SMOTE, Random Oversampling (ROS), and Random Undersampling (RUS). A Novel class imbalance processing technology, referred to as SGM-CNN [58], has been proposed that combines SMOTE and undersampling for clustering based on Gaussian Mixture Model (GMM). They then integrate imbalanced class processing with a convolutional neural network to design a flow-based intrusion detection model. Their experimental results show that SGM-CNN provides an effective solution to imbalanced intrusion detection.

This study [57] has proposed a hybrid filter-wrapper feature selection algorithm that selects robust features, i.e. features that are resistant to concept drift and represent minority classes. Experimental results over Cambridge Intrusion Dataset [40] show that proper feature selection leads to higher classification accuracy and better F-measure for each class, especially minority classes. They provide a wide variety of features to characterize flows that includes simple statistics about packet length and inter-packet timings as well as information from the transport protocol. Cost-sensitive learning of deep neural networks is also widely studied [9], [26], [46]. A Flexible Neural Tree (FNT) can search optimal network structures using tree structure evolving algorithms which results in high performance in predictive modeling. A new similarity evaluation method for FNT has been proposed [45] to keep the population diversity and deal with imbalanced data. Their proposed method uses an imbalanced fitness function to control its evolving procedure to deal with imbalanced data problems.

A survey of recent literature on the class imbalance problem in the context of deep learning [18] suggests that while increasing the depth of neural networks is beneficial to their robustness and predictive power, depth alone is not sufficient to deal with the problem of class imbalance in the cases of MLP or Convolutional Neural Networks (CNNs). Their experiments on CNNs show that too deep a network may, in fact, be harmful. The same study [18] finds that while regularization helps improve classification performance in some domains, these improvements remain insignificant in the context of class imbalance.

A systematic study on CNNs [5] finds oversampling to be the most effective method in addressing the class imbalance problem next to cluster-based oversampling, SMOTE, and RUS over three benchmark datasets MNIST, CIFAR-10 and ImageNet. They assert that it does not cause overfitting of CNNs and suggest applying it to the level that completely eradicates the imbalance. They conclude that undersampling performs on par with oversampling in the case of extreme ratios of imbalance and most classes being in the minority.

Given that network traffic incorporates timestamps, this study [49] adopts phased processing prior to classification, i.e. they deploy a Bidirectional Long Short-term Memory (BLSTM) model to learn the sequential features in the traffic data. Next, the attention layer is used for feature learning on the sequential data.

In [38], we present a preliminary study on learning from SDN intrusion data [12] where SMOTE remarkably improves classification results. In the current paper, we present a comprehensive study of the class imbalance problem in SDN intrusion data and extend the work in [37] significantly by investigating the benefits of multiple balancing strategies along with imbalanced learning techniques that acknowledge the class imbalance problem and incorporate appropriate strategies in the learning process so as to improve classification performance in minority classes where a higher detection rate is of importance. Although existing literature has extensively covered the class imbalance problem, addressing the class imbalance problem in machine learning-based intrusion detection in SDN remains to be a research gap. This paper provides comprehensive benchmarking and a detailed discussion of the class imbalance problem in SDN intrusion data.

III. METHODS

Data-level methods require a balancing phase prior to classification whereas classifier-level methods do not call for balancing and directly incorporate imbalanced learning within the learning phase. Fig. 1 demonstrates the methods used in this study to learn from the imbalanced dataset. We selected balancing methods from two main categories: data-level and classifier-level. Our study applies four data-level balancing methods, RUS, ROS, SMOTE, and GANs, to three learning models (deep learning and Decision Trees): (1) MLP with 6 layers (model MLP1) (2) MLP with 10 layers (model MLP2), (3) Random Forest (model RF with hyperparameter max_depth=10). We selected two classifier-level methods: weighted RF and one-shot learning via Siamese Neural Networks. Here, we present the details of the balancing strategies and utilized learning models.

A. DATASET

To evaluate the proposed approach, we use the InSDN dataset that provides 343889 network flows captured within a synthetic SDN testbed. In the evaluation, we compare sampling techniques against three baseline classifiers, namely MLP and RF learning models. Then, we compare imbalanced learning approaches separately as they do not require a dataset balancing phase prior to classification.

B. DATA PRE-PROCESSING

Network Identifiers such as source IP, Destination IP, and flow ID are removed in order to avoid the overfitting problem on account of the fact that they can be changed from network to network. Moreover, 8 zero variation features in the InSDN dataset that do not contain any information useful for classification are eliminated. Next, features are standardized to restrict the scale of the values between -1 and 1.

TABLE 1. Distribution of samples over data-level methods.

Method	Class	Train	Test
pre-balancing	Normal	47897	20527
	DDoS	85359	36583
	DoS	37531	16085
	Probe	68690	29439
	BFA	984	421
	Web-Attack	134	58
	Botnet	115	49
	U2R	12	5
ROS	Normal	85359	20527
	DDoS	85359	36583
	DoS	85359	16085
	Probe	85359	29439
	BFA	85359	421
	Web-Attack	85359	58
	Botnet	85359	49
	U2R	85359	5
SMOTE	Normal	69236	20527
	DDoS	85359	36583
	DoS	58870	16085
	Probe	111369	29439
	BFA	43663	421
	Web-Attack	42813	58
	Botnet	42794	49
	U2R	42691	5
GANs	Normal	58137	20527
	DDoS	86639	36583
	DoS	42651	16085
	Probe	72530	29439
	BFA	13784	421
	Web-Attack	2694	58
	Botnet	10355	49
	U2R	10252	5

C. EVALUATION

Our evaluation metrics for classification performance results are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions. Measures of accuracy, precision, recall, and F1-score [44] are computed

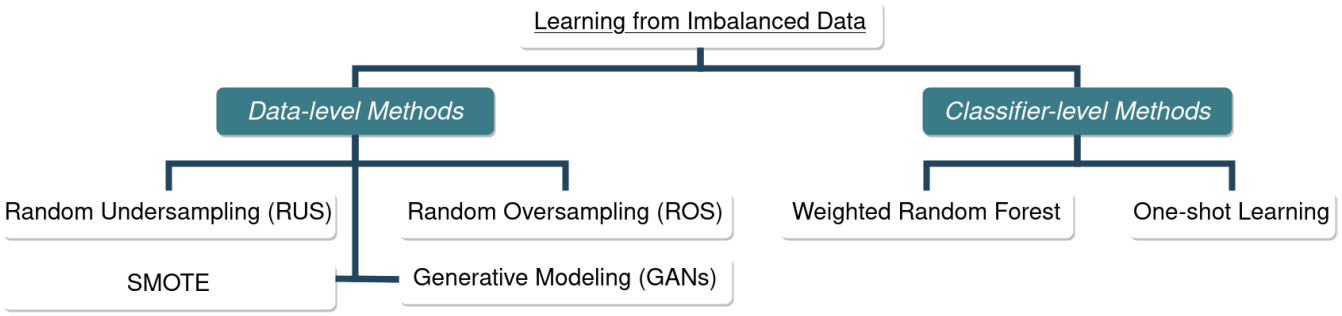


FIGURE 1. Data-level and classifier-level approaches to learning from imbalanced data.

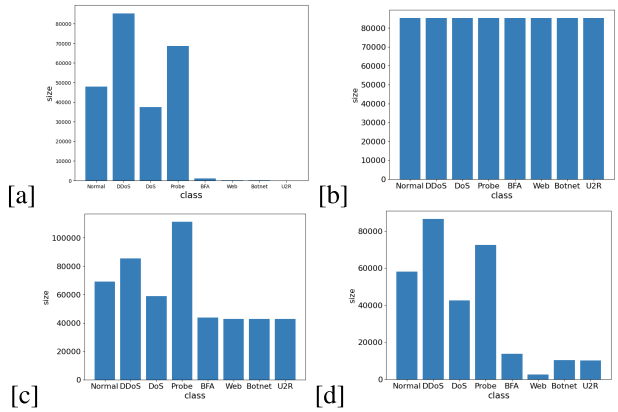


FIGURE 2. Distribution of network traffic samples across classes with data-level methods: (a) Pre-balancing (b) ROS (c) SMOTE (d) GANs.

directly from the abovementioned metrics for assessing the performance of each intrusion type in a multi-class model. Measures of Macro and Micro F1 are used for the cumulative evaluation of a model for all intrusion types. Macro F1 better reflects the classification performance over imbalanced data as it considers each intrusion type equally regardless of the representation of that type in the datasets. Macro F1 better reflects a model's classification performance in minority classes. 1), precision (equation 2), recall (equation 3) and F1 score (equation 6) are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Macro Average Precision (equation 4) and Macro Average Recall (equation 5) are calculated as the arithmetic mean of the Precision and Recall values for each class [20].

$$P = MacroAveragePrecision = \frac{\sum_{k=1}^{K} Precision_k}{k} \tag{4}$$

$$R = MacroAverageRecall = \frac{\sum_{k=1}^{K} Recall_k}{k} \tag{5}$$

Macro F1-Score is computed as the harmonic mean of Macro-Precision and Macro-Recall [20] as given below:

$$MacroF1score = 2 * \frac{P * R}{P^{-1} + R^{-1}} \tag{6}$$

Finally, Micro F1-Score is computed from Micro-Precision and Micro-Recall, which are weighted averages of Precision and Recall values of all classes.

We carried out the experiments on a processor with 2199.998 MHz, 13G memory, and 56320K L3 cache. Experiments were implemented using Python and PyTorch [43] library.

D. DATA-LEVEL METHODS

An imbalanced dataset is one where classes are not approximately equally represented. Imbalanced classification poses a challenge for predictive modeling as most machine learning algorithms were designed around the assumption of an equal number of samples for each class. Learning systems trained with imbalanced data usually fail to recognize minority class instances effectively. InSDN [12] consists of 7 distinct intrusion types. The majority of traffic flows belong to *Normal*, *DDoS*, *DoS*, and *Probe* classes with the remaining classes (*BFA*, *Web-Attack*, and *U2R*) making up about 1% of the entire dataset. Therefore, a learning system may not effectively learn decision boundaries for minority classes which are important to identify.

The goal of sampling methods in the context of imbalanced learning is to modify an imbalanced dataset through a mechanism where a balanced distribution is provided among classes to enhance overall classification performance [13], [21], [28], [55].

1) RANDOM OVERSAMPLING (ROS) AND RANDOM UNDERSAMPLING (RUS)

In ROS, the original set is augmented by replicating randomly selected minority examples, increasing the number of total examples. In RUS, randomly selected majority class examples are removed from the dataset, adjusting the balance of the original dataset. While oversampling and undersampling provide balance to the dataset, they introduce their own set of problematic consequences in the context of learning systems. For example, removing examples from

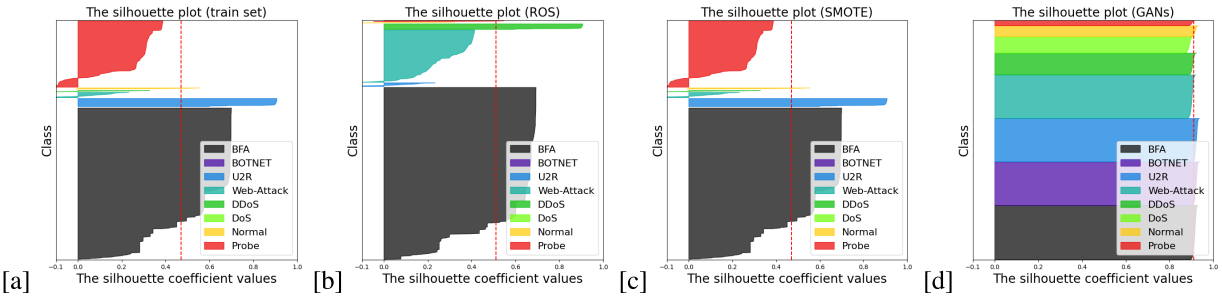


FIGURE 3. The silhouette plot for data-level methods: (a) train set (average silhouette score: 0.46) (b) ROS (average silhouette score: 0.46) (c) SMOTE (average silhouette score: 0.51) (d) GANs (average silhouette score: 0.91). The silhouette plot shows a measure of the compactness of each class, i.e. how close each point in one class is to the points in the neighboring classes. We observe that while ROS improves the silhoutte score, SMOTE retains the same score. SMOTE seems to retain the structure of classes more accurately whereas GANs seem to improve balance among classes.

TABLE 2. Optimal set of hyperparameters for DNN models.

Model	HL	HU	LR	Optimizer	AF	Batch-Size	Epoch
MLP1	6	52, 64, 128, 64, 16, 8	0.0001	Adam	ReLU	1024	200
MLP2	10	52, 128, 512, 1024, 1024, 512, 256, 128, 64, 8	0.0001	Adam	ReLU	1024	200

TABLE 3. Optimal set of hyperparameters for GANs models.

Model	Convolutional Layers	Kernels	LR	Optimizer	AF	Batch-Size	Epoch
Discriminator	64*, 128, 256, 1	(4,4)*, (3,3)	0.0002	Adam	LeakyReLU	64	200
Generator	256*, 128, 64, 1	(3,3)*, (4,4)	0.0002	Adam	ReLU	256	200

TABLE 4. Optimal set of hyperparameters for siamese neural network.

Model	Kernels	Optimizer	Batch-Size	Epoch
Siamese Neural Networks	(7,7), (3,3)	AdaDelta	64	200

the majority class may cause the learning system to miss important information and replicating instances may lead to overfitting [10].

2) SMOTE

In this method, minority classes are over-sampled by creating synthetic examples rather than over-sampling with replacement. In this fashion, the oversampling of minority classes is achieved by taking each sample and introducing synthetic examples along the line segments joining any/all of the k nearest neighbors [8]. Authors of SMOTE [8] suggest combining it with random undersampling of the majority class. These synthetic examples lead to creating larger and more specific decision regions in the model, resulting in better generalization of learning algorithms where more general regions are learned for the minority class examples rather than by the majority class examples around them [8].

3) GENERATIVE MODELING

The core idea behind generative modeling in the context of tackling the class imbalance problem is to estimate the probability density function describing the data and generate new data instances in a random fashion [39] in order to balance the data distribution in an otherwise imbalanced dataset. Generative models typically construct a latent space that aims to capture the direct cause of the target variable. The latent space is represented by a probability distribution over possible values rather than a single fixed value, enforcing uncertainty onto the model, which may lead to more stable predictions. Randomness is also incorporated into the data generation process so that we can arrive at well-justified sampling variability considerations from a statistical point of view. The idea behind GANs is to create a contrived game between two deep learning models, namely the discriminator and the generator networks [19]. The generator model continues to synthesize data points that highly resemble the original data distribution, whereas the discriminator model continues to evaluate whether the synthetic data point actually belongs to the original distribution. GANs are based on game theory, while most other approaches to generative modeling are based on optimization [19]. The goal of the game is for the generator to synthesize data points that the discriminator believes to belong to the original distribution. In this way, realistic synthetic data are generated for the task of data augmentation. Fig. 4 shows the architecture and dimensions of the proposed conditional Deep Convolutional Generative Adversarial Networks model. Table 3 shows the optimal hyperparameters for our GANs model. Network flow instances are reshaped into image-like structures prior to entering the discriminator network. Both the generator and discriminator are CNNs.

E. CLASSIFER-LEVEL METHODS

Although data-level methods explained above are widely used, this does not, however, imply that classifiers cannot learn directly from imbalanced datasets. In fact, it is

TABLE 5. Hyperparameters for random forests.

Model	Estimators	Classes	Features	Maximum Depth	Minimum Samples in a Leaf	Minimum Split Samples
RF, wRF	100	8	52	10	1	2

TABLE 6. Ranges of hyperparameters for deep learning models.

Model	LR	Batch-size	Epoch	Layers
MLP1	0.0001-0.0002	512-1024	100-2000	52,64,128,512,512,128,64,16,8 52,64,128,512,128,64,16,8
MLP2	0.0001-0.0002	512-1024	100-2000	52,128,512,1024,1024,512,256,128,64,8 52,128,128,512,1024,1024,512,256,128,128,64,8
Discriminator (GANs)	0.0001-0.0002	64-128	100-2000	64,128,256,1 128,256,512,128,1
Generator (GANs)	0.0001-0.0002	64-256	100-2000	256,128,64,1 64,128,256,128,1
Siamese Neural Networks	0.01-0.0001	32-256	100-2000	64,64,128,256,256,512,512,1

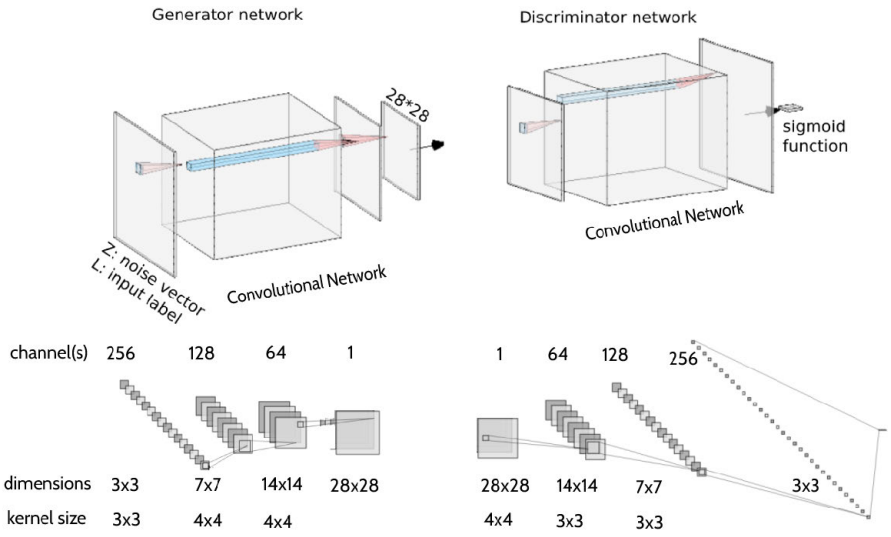


FIGURE 4. Architecture of cDCGANs framework.

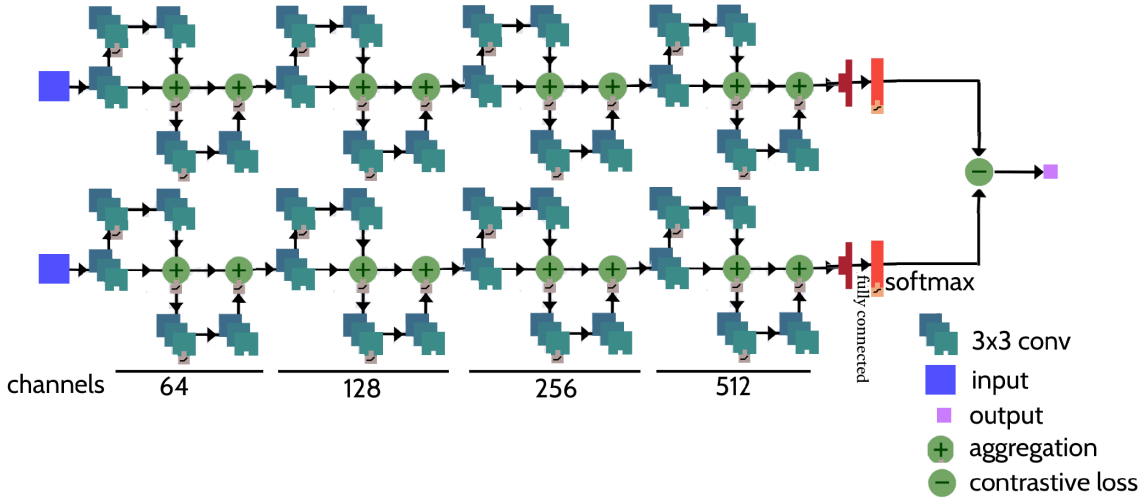


FIGURE 5. Architecture of siamese neural networks framework.

shown that classifiers trained on some imbalanced datasets exhibit comparable performance metrics to those trained on data sets balanced by sampling techniques [4], [24]. Here, we discuss distance-based learning using Siamese neural networks and weighted Random Forest as effective methods to learn from imbalanced data distributions. Tables 5 and 6 describe the hyperparameters of baseline classifiers.

1) BASELINE CLASSIFIERS

To detect intrusions in InSDN network traffic data, a multi-class classification approach was adopted where the dataset was split into train and test sets by a 70%-30% train-test ratio while preserving the stratified ratio of all classes. Network flows were reshaped to image-like structures (28*28 matrices) to fit the dimensions of the convolutional layers within the proposed GANs and Siamese Neural Networks models. The following baseline classifiers were selected:

- MLP with 6 layers (referred to as *MLP1*) with 52, 64, 128, 64, 16, and 8 nodes in each layer, respectively;
- MLP with 10 layers (referred to as *MLP2*) with 52, 128, 512, 1024, 1024, 512, 256, 128, 64, and 8 nodes in each layer, respectively.
- RF: We use a Random Forest Classifier with a maximum depth of 10. Class weights are not specified, assigning the same weight to all classes. The number of estimators in Random Forest classification is 100.

2) ONE-SHOT LEARNING

Learning from only a few examples remains a key challenge in machine learning. One-shot learning has emerged successful in computer vision [27], where objects are classified after learning from only a few examples. In this method, the classification of images is performed based on their similarity, not the analysis of a large number of features.

One-shot learning can be achieved through different methods using Siamese networks to evaluate the probability that input pairs belong to the same class. Siamese Networks are made up of twin branches of identical neural networks whose outputs are used to learn the contrastive loss function between data pairs, also referred to as the distance between them. These twin CNNs share weights in the training process of data pairs. The resulting feature space, in fact, represents distances between data points in the latent feature space. A number of studies such as [33] and [34] have utilized Siamese networks with active learning in the non-stationary data stream. One-shot learning offers excellent feature extraction capabilities in distance-based learning [14]. In this fashion, the similarity between any two data points can be measured to determine whether they belong to the same class.

A Similarity-based Intrusion Detection [23] has been proposed that leverages the Siamese architecture and a majority voting scheme for classification provided that the generation of data pairs during the training process complies with the following constraints [23]: (a) uniqueness of data pairs and (b) balanced representation of all combinations. Fig. 5 shows the architecture of the proposed Siamese Neural Networks model. It is a variant of the ResNet [22] architecture. Deep Residual Learning [22] explicitly reformulates the layers as learning residual functions with reference to the layer inputs instead of learning unreferenced functions. The bottleneck design is used to prevent high-time complexity when the network is very deep. Moreover, skip connections within the network help resolve the vanishing gradient problem. Table 4 shows the optimal set of hyperparameters for the Siamese Neural Networks model.

3) WEIGHTED RANDOM FOREST

The traditional RF method utilizes an ensemble of classification trees to predict the outcome from predictors, with each tree trained on a different sample of N subjects, and random subset predictors considered at each node of the tree. RF then aggregates tree-level results equally across trees. wRF [56] utilizes performance-based weights for tree aggregation where votes from each tree in the forest are considered in such a fashion that better-performing trees are weighted more heavily. Weighted Random Forest can easily handle imbalanced data by forming ensembles with weights for different classes. Therefore, it does not require dataset balancing prior to classification.

IV. RESULTS

For the purpose of comparison, the performance of learning directly from imbalanced data is given as a baseline. Besides, we compare the proposed approach with the fundamental methods for processing imbalanced data, .i.e., ROS and SMOTE. RUS was used only alongside SMOTE as recommended by its authors [8].

A. DATA-LEVEL METHODS

InSDN incorporates a high degree of imbalance between classes (Fig. 2 (a)) where minority classes, BFA, Web, Botnet, and U2R, make up less than 1% of the entire dataset. We balance class distributions prior to classification via data-level methods. Fig. 2 shows the distribution of classes before and after data augmentation using ROS, SMOTE, and GANs. Table 1 reports the exact sizes of the train and test sets with respect to each method.

As reported in Table 7, our baseline classifiers (i.e., pre-balancing row) offer great performance over majority classes based on precision, recall, and F1-score values (99% and above) for Normal, DDoS, DoS, and Probe classes. However, all three classifiers (MLP1, MLP2, and RF) perform poorly over minority classes. For example, deep learning models offer a 50% F1-score in the detection of Web-Attack instances at best and a 0% in U2R. Moreover, while Random Forest offers adequate performance in the classification of minority classes (F1-score values above 74%), it suffers from high numbers of False Negatives (FN), i.e. lower recall values. It can be observed that Random Forest's classification performance with respect to U2R (60% Precision) and BFA (81.23% Precision) is relatively lower than other classes. We observe that Random Forest has performed well over imbalanced data, i.e. RF's F1-score for classification over imbalanced dataset is the best compared to other results from data-level methods with the exception of U2R and Botnet classes in which case RF's F1-score is close to the best classification performance through SMOTE and ROS, respectively.

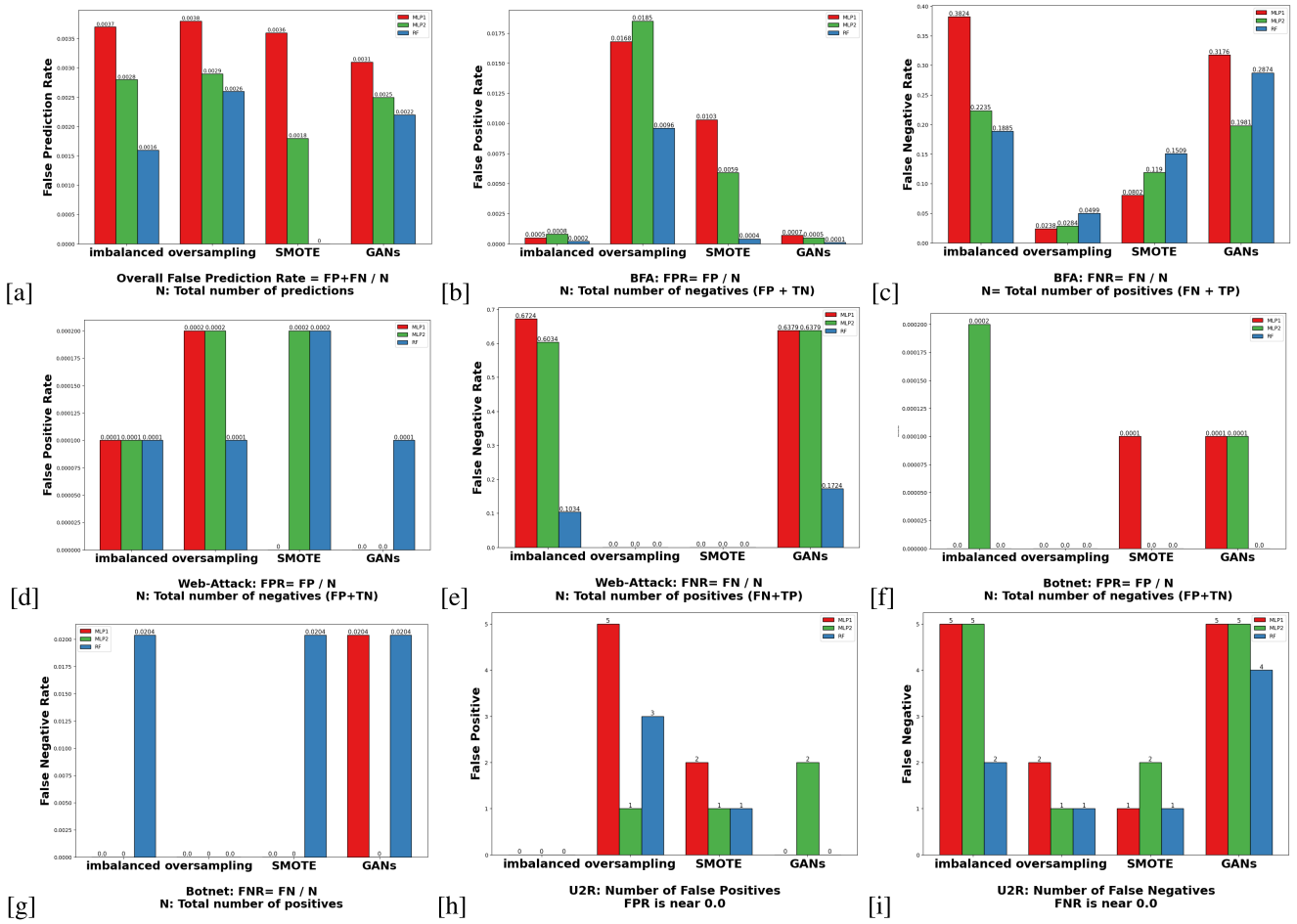


FIGURE 6. False Prediction Rates of minority classes in InSDN dataset, i.e. BFA, Web-Attack, Botnet, and U2R with different sampling methods: (a) Overall false prediction rate (b) BFA: False positive rate (c) BFA: False negative rate (d) Web-Attack: False positive rate (e) Web-Attack: false negative rate (f) Botnet: False positive rate (g) Botnet: False Negative rate (h) U2R: Number of false positives (i) U2R: Number of false negatives.

An important evaluation metric regarding classification over imbalanced data is the difference between Micro and Macro F1-score values which indicates a classifier's behavior on imbalanced datasets. Micro F1-score does not usually reflect an objective measure of classification performance when the classes are imbalanced, while Macro F1-score does in fact reflect the class imbalance problem in classification accuracy. We can see in Table 9 that while Micro F1-score values are remarkable for all baseline classifiers, Macro F1-score says more about a classifier's ability to deal with imbalanced data. For example, while pre-balancing Micro F1-score values for MLP1 and MLP2 are above 99%, Macro F1-score values are only between 76 % and 78 %. After balancing via SMOTE, however, we can see that Macro F1-score values actually increased for MLP1 and MLP2 as Micro F1-score slightly decreased. When we explore the specific changes in classes as given in Table 7, SMOTE slightly decreases the F1 performance of the majority class, Probe, from 99.71% to 97.97% in MLP1. Although this balancing strategy increases the F1 for U2R class remarkably from 0% to 72.72%, it degrades the result of the other minority class, BFA. These results indicate that balancing strategies may have varying effects on majority and minority classes, requiring special attention.

Here, we compare results from classification with ROS, SMOTE, and GANs. Table 7 shows classification performance with data-level methods in terms of accuracy, precision, recall, and F1-score per class. Experimental results reported in Table 9 give an overall view of classification performance, which suggests that while ROS, SMOTE, and GANs all offer improvement to classification performance in MLP1 and MLP2, SMOTE has yielded better overall Macro-F1 values than ROS and GANs. However, the contribution of GANs is very limited for deep learning models. On the other side, ROS and GANs decreased the Macro-F1 performance of RF. We obtained similar, excellent performance from all three baseline classifiers for the majority classes, i.e., Normal, DDoS, DoS, and Probe. An interesting finding is that RF outperforms both MLP1 and MLP2 models in the classification of all minority classes. This suggests that Random Forest can handle class imbalance well. In fact, it can be observed that balancing strategies result in the degradation of RF's classification performance over minority classes, with the exception of a slight enhancement achieved through

TABLE 7. Classification performance with data-level methods.

	Normal			DDoS			DoS			Probe		
Classifier:	MLP1	MLP2	RF	MLP1	MLP2	RF	MLP1	MLP2	RF	MLP1	MLP2	RF
Accuracy												
Pre-balancing	99.86 %	99.93 %	99.97 %	99.98 %	99.98 %	99.99 %	99.84 %	99.89 %	99.94 %	99.83 %	99.84 %	99.87 %
ROS	99.92 %	99.95 %	99.95 %	99.98 %	99.99 %	99.99 %	99.89 %	99.92 %	99.90 %	98.23 %	98.08 %	98.98 %
SMOTE	99.88 %	99.97 %	99.96 %	99.99 %	99.99 %	99.99 %	99.84 %	99.94 %	99.93 %	98.86 %	99.33 %	99.85 %
GANS	99.93 %	99.94 %	99.96 %	99.99 %	99.98 %	99.99 %	99.85 %	99.91 %	99.88 %	99.85 %	99.84 %	99.85 %
Recall												
Pre-balancing	99.72 %	99.83 %	99.97 %	99.98 %	99.98 %	99.97 %	99.78 %	99.65 %	99.83 %	99.74 %	99.78 %	99.85 %
ROS	99.73 %	99.88 %	99.85 %	99.96 %	99.98 %	99.99 %	99.70 %	99.76 %	99.81 %	93.94 %	93.35 %	96.50 %
SMOTE	99.53 %	99.91 %	99.97 %	99.98 %	99.99 %	99.99 %	99.57 %	99.76 %	99.79 %	96.21 %	97.89 %	99.72 %
GANS	99.81 %	99.89 %	99.95 %	99.99 %	99.97 %	99.99 %	99.77 %	99.72 %	99.80 %	99.76 %	99.79 %	99.81 %
Precision												
Pre-balancing	99.58 %	99.86 %	99.92 %	99.98 %	99.97 %	100 %	99.21 %	99.65 %	99.78 %	99.69 %	99.68 %	99.69 %
ROS	99.90 %	99.90 %	99.91 %	99.98 %	99.99 %	100 %	99.60 %	99.73 %	99.56 %	99.86 %	99.91 %	99.92 %
SMOTE	99.88 %	99.95 %	99.86 %	99.99 %	100 %	100 %	99.41 %	99.87 %	99.81 %	99.80 %	99.76 %	99.75 %
GANS	99.86 %	99.83 %	99.87 %	99.98 %	99.98 %	99.99 %	99.31 %	99.72 %	99.45 %	99.71 %	99.65 %	99.68 %
F1-score												
Pre-balancing	99.65 %	99.84 %	99.94 %	99.98 %	99.97 %	99.99 %	99.49 %	99.65 %	99.81 %	99.71 %	99.73 %	99.77 %
ROS	99.81 %	99.89 %	99.88 %	99.97 %	99.99 %	99.99 %	99.65 %	99.74 %	99.69 %	96.81 %	96.52 %	98.18 %
SMOTE	99.70 %	99.93 %	99.91 %	99.99 %	99.99 %	99.99 %	99.49 %	99.82 %	99.80 %	97.97 %	98.82 %	99.74 %
GANS	99.84 %	99.86 %	99.91 %	99.98 %	99.97 %	99.99 %	99.54 %	99.72 %	99.63 %	99.74 %	99.72 %	99.75 %

	BFA			Web-Attack			BOTNET			U2R		
Classifier:	MLP1	MLP2	RF	MLP1	MLP2	RF	MLP1	MLP2	RF	MLP1	MLP2	RF
Accuracy												
Pre-balancing	99.78 %	99.83 %	99.90 %	99.95 %	99.98 %	99.98 %	99.99 %	99.98 %	99.99 %	99.99 %	99.99 %	99.99 %
ROS	98.29 %	98.12 %	99.02 %	99.97 %	99.98 %	99.98 %	99.99 %	100 %	99.99 %	99.99 %	99.99 %	99.99 %
SMOTE	98.89 %	99.36 %	99.90 %	99.97 %	99.98 %	99.98 %	99.99 %	100 %	99.99 %	99.99 %	99.99 %	99.99 %
GANS	99.79 %	99.86 %	99.86 %	99.96 %	99.95 %	99.98 %	99.98 %	99.98 %	99.99 %	99.99 %	99.99 %	99.99 %
Recall												
Pre-balancing	61.75 %	77.43 %	81.23 %	32.75 %	39.65 %	89.65 %	100 %	100 %	97.95 %	0 %	0 %	60 %
ROS	97.62 %	97.14 %	95.01 %	100 %	100 %	100 %	100 %	100 %	100 %	60 %	80 %	80 %
SMOTE	91.92 %	88.12 %	84.79 %	100 %	100 %	100 %	100 %	100 %	97.95 %	80 %	60 %	80 %
GANS	67.93 %	80.04 %	71.25 %	36.20 %	36.20 %	82.75 %	97.95 %	100 %	97.95 %	0 %	0 %	20 %
Precision												
Pre-balancing	82.27 %	80.29 %	95.53 %	65.51 %	67.64 %	89.65 %	90.74 %	74.24 %	97.95 %	0 %	0 %	100 %
ROS	18.99 %	17.51%	28.92 %	69.87 %	75.32 %	79.45%	96.07 %	100 %	98 %	37.5 %	80 %	57.14 %
SMOTE	26 %	37.97 %	90.15 %	65.90 %	74.35 %	78.37 %	87.5 %	100 %	97.95 %	66.66 %	75 %	80 %
GANS	79 %	86.18 %	95.23 %	87.5 %	80.76 %	88.88 %	76.19 %	79.03 %	96 %	0 %	0 %	100 %
F1-score												
Pre-balancing	70.55%	78.83 %	87.80 %	43.67 %	50%	89.65 %	95.14 %	85.21 %	97.95 %	0 %	0 %	74.99 %
ROS	31.79 %	29.68%	44.34 %	82.26 %	85.92 %	88.54 %	98 %	100 %	98.98 %	46.15 %	80 %	66.66 %
SMOTE	40.54 %	53.07 %	87.39 %	79.45 %	85.29 %	87.87 %	93.33 %	100 %	97.95 %	72.72 %	66.66 %	80 %
GANS	73.05 %	83 %	81.52 %	51.21 %	50 %	85.71 %	85.71 %	88.28 %	96.96 %	0 %	0 %	33.33 %

TABLE 8. Classification performance with imbalanced learning.

	Normal		DDoS		DoS		Probe	
Classifier:	SIM	wRF	SIM	wRF	SIM	wRF	SIM	wRF
Accuracy	99.84 %	99.99 %	99.96 %	99.99 %	96.02 %	99.97 %	85.74 %	99.86 %
Precision	99.63 %	99.98 %	99.96 %	99.99 %	89.39 %	99.88 %	98.15 %	99.79 %
Recall	99.59 %	99.96 %	99.94 %	100 %	84.56 %	99.92 %	51 %	99.73 %
F1-score	99.61 %	99.97 %	99.95 %	99.99 %	86.91 %	99.90 %	67.12 %	99.76 %

	BFA		Web-Attack		BOTNET		U2R	
Classifier:	SIM	wRF	SIM	wRF	SIM	wRF	SIM	wRF
Accuracy	97.45%	99.89 %	99.90%	99.99 %	99.99 %	99.99 %	87.88	99.99%
Precision	11.41%	84.56 %	35.29%	98.27 %	87.5%	97.95 %	0 %	60 %
Recall	77.43%	88.55 %	93.10%	89.06 %	100 %	97.95 %	0 %	100 %
F1-score	19.89 %	86.51 %	51.18 %	93.44 %	93.33 %	97.95 %	0 %	74.99 %

wRF: weighted Random Forest, SIM: Similarity-based Classification with Siamese Neural Networks

oversampling over Botnet (a minority class in InSDN). GAN-based data augmentation has proven to be helpful in better classification of BFA, whereas statistical methods such as ROS and SMOTE improved the classification of Web-Attack and U2R. Fig. 7 shows confusion matrices for classification before and after data augmentation using ROS, SMOTE, and GANs. The numbers inside each confusion matrix add up to 103, 167 flows (the size of our test set). All confusion

TABLE 9. Macro-F1 and Micro-F1 values for different approaches.

	Pre-balancing		Oversampling		SMOTE		GANs	
Metric:	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
MLP1	76.82 %	99.63 %	85.03 %	98.14 %	87.61 %	98.72 %	77.60 %	99.68 %
MLP2	77.35 %	99.71 %	89.74 %	98.02 %	89.38 %	99.29 %	78.75 %	99.74 %
RF	94.32 %	99.84 %	89.12 %	98.92 %	94.25%	99.81 %	90.17 %	99.77 %
wRF	94.67 %	99.85 %						
SIM	70.04 %	83.41 %						

wRF: weighted Random Forest, SIM: Similarity-based Classification with Siamese Neural Networks

matrices were created from the predicted data of the trained model over the test set. The values of the main diagonal correspond to the values correctly estimated by the model. Evaluating classification performance over different baseline classifiers such as MLPs and RF also helps understand how these classifiers perform over data that is balanced via different balancing strategies. Fig. 6 shows the number of false predictions with each balancing method over baseline classifiers for minority classes, i.e. BFA, Web-Attack, Botnet, and U2R. For example, we observe that ROS and SMOTE drastically decrease the False Positive (FP) rate but increase the False Negative (FN) in the classification of BFA, Web-Attack, and U2R. Fig. 6 (a) shows that while data-level methods usually improve classification performance over deep learning models, i.e. reduce the overall number of false predictions, they, in fact, degrade RF's performance, i.e. cause higher numbers of false predictions.

B. CLASSIFIER-LEVEL METHODS

Table 8 reports the performance metrics of imbalanced learning using wRF and Similarity-based classification (SIM) methods. Both methods offer remarkable performance in the classification of Normal and DDoS instances with measures of precision, recall, and F1-score above 99%. Similarity-based classification (SIM) of DoS and Probe classes, however, exhibits lower overall accuracy than wRF. Only 51% of instances detected as Probe by SIM were in fact the Probe class. SIM offers very low performance in the detection of minority classes, especially BFA and U2R where it achieves a precision of 11.41% and 0%, respectively. While wRF outperforms SIM in every class, SIM has detected Botnet instances with comparable confidence and an even better recall measure. Moreover, all instances detected as Botnet by SIM were in fact correctly classified, whereas wRF has a non-zero False Positive (FP) rate. However, the Botnet class tends to be easier to detect than other minority classes, for example, according to Table 7 data-level methods achieve the highest F1 scores for this particular minority class. Fig. 8 shows confusion matrices classifier-level methods.

V. DISCUSSION

We observe that many studies do not pay particular attention to identifying metrics suitable for imbalanced datasets. One of the strengths of this study is our metric selection for multi-class classification, where we present Macro and Micro F1-score values as well as per-class measures of accuracy, precision, recall, and F1. Although Micro-F1 is useful in understanding the impact of balancing strategies on the entire dataset, including the majority ones, Macro-F1 is more instrumental in demonstrating the performance of the model in minority ones as it treats each class equally. Therefore, a lower score in a minority class is easily reflected in the overall score. Nevertheless, our study also presents accuracy, precision, recall, and F1 for each class, enabling us to comprehend the balancing effect on a per-class basis.

Learning algorithms handle the class imbalance problem in different ways. In Random Forest, for example, Small changes to training data can result in a significantly different tree structures. Perhaps, one reason why Random Forest handles class imbalance fairly well is that Random Forest is an ensemble method combining results from multiple decision trees. As reported in Table 9, wRF performs well in both Macro and Micro F1-scores which is indicative of the effectiveness of weighted ensembles within the wRF's tree structure. RF and wRF have similar results demonstrating the effectiveness of RF in the face of data imbalance. It is interesting to observe that an RF model without any balancing strategy outperforms all complex deep learning models with varying balancing strategies according to Macro-F1 scores.

It is observable that balancing strategies such as ROS, SMOTE, and GANs all offer improvement to the classification performance of baseline deep learning models, MLP1 and MLP2. However, GANs induce minuscule improvement in the model performance. Even though GANs have proven successful in generating realistic synthetic data in multiple domains, such as Computer Vision and Natural Language Processing, this generative modeling approach has offered little improvement in the case of InSDN. SMOTE has outperformed ROS and GANs by offering the utmost overall improvement over the classification results of deep learning models. With respect to baseline classifiers, we can observe that Random Forest performs well without any data balancing prior to classification. Interestingly, ROS and GANs results decrease Macro-F1 performance, whereas SMOTE results remain almost the same.

Our findings reaffirm the findings of other studies [15], [29] on the class imbalance problem that suggests ensemble-based methods tend to produce good classification results over imbalanced datasets, usually better than data-level methods. A study on the class imbalance problem in

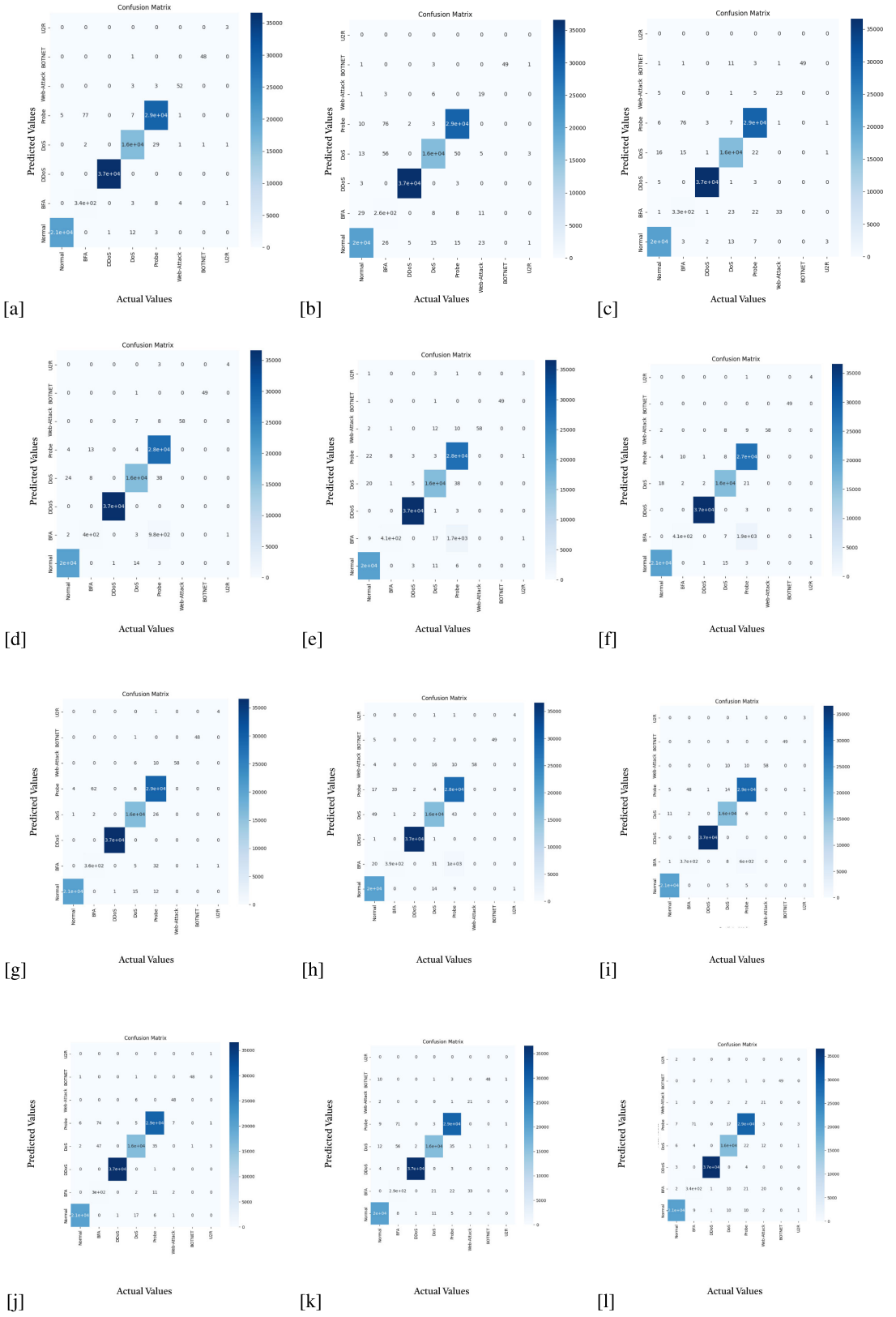
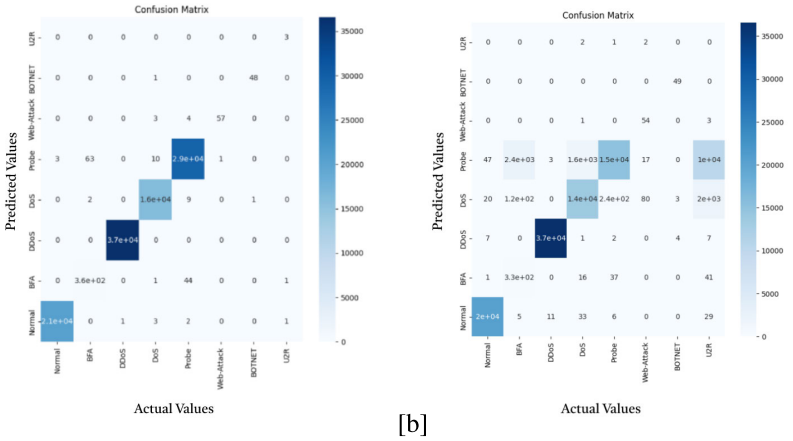


FIGURE 7. Confusion matrices for classification before and after data augmentation, N = 103167 (a) RF: Pre-balancing (b) MLP1: Pre-balancing (c) MLP2: Pre-balancing (d) RF: Oversampling (e) MLP1: Oversampling (f) MLP2: Oversampling (g) RF: SMOTE (h) MLP1: SMOTE (i) MLP2: SMOTE (j) RF: GANs (k) MLP1: GANs (l) MLP2: GANs.



[a] [b]

FIGURE 8. Confusion matrices for imbalanced learning (a) wRF (b) SIM.

classical NSL-KDD dataset [47] similarly finds that Random Forest outperforms other classifiers. None of the classifiers provides adequate performance in the case of U2R, which is the smallest minority class. It is a common finding that ensemble-based classifiers work well for minority classes, however, for an extremely rare class such as U2R, even Random Forest fails to perform adequately. Even though SMOTE has been considered [48] a suboptimal alternative for research problems in high-dimensional space because of its orientation towards feature space rather than data space, we have found that SMOTE has outperformed more sophisticated data augmentation methods such as GANs.

We covered data-level and classifier-level methods for addressing the class imbalance problem in SDN intrusion data. Here, we summarize the benefits of each technique with respect to each minority class as supported by our experiments. With deep learning baseline classifiers, we observe that

- GANs improved the detection of BFA significantly.
- ROS improved the detection of Web-Attack and Botnet classes.
- ROS and SMOTE improved the detection of U2R.

Due to the ensemble nature of random forest classifier, it can handle class imbalance well. With random forest classifier, we observe that while ROS improved the detection of Botnet and U2R, random forest performed well with no data-level methods in the detection of BFA and Web-Attack classes.

In discussing the limitations of our study, we hypothesize that contrastive learning via Siamese Networks may require hefty fine-tuning of the convolutional neural networks as well as the data pairing process during training. We reckon pairing Siamese networks with other classifiers would probably produce more robust results compared to the similarity-based learning approach. In this fashion, Siamese networks would be utilized as a feature extractor. The similarity-based detection approach proposed in [23] seems to not be able to differentiate between the minority classes, which are all the more important to detect.

VI. CONCLUSION AND FUTURE WORK

This paper investigates the class imbalance problem in machine learning-based intrusion detection in SDN. It has been shown that the detection rate can be severely affected by the imbalanced distribution of attack classes which is a widespread issue with most intrusion datasets. Data balancing strategies (data-level methods) such as Random Oversampling (ROS), SMOTE, and generative modeling via Generative Adversarial Networks (GANs) have been proposed to mitigate the impact of class imbalance on classification performance. Moreover, imbalanced learning methods (classifier-level approach) such as weighted Random Forest and Siamese Neural Networks for one-shot learning have been adopted which render promising in the detection of minority class instances without the need for a balancing phase prior to classification. In particular, weighted RF yielded the best performance throughout the experiments described in the previous section. However, the original RF provided almost the same performance. This is indicated by weighted RF's only slightly superior Macro and Micro-F1 values of 99.85% and 94.67% as compared to RF's Macro and Micro-F1 values of 99.84% and 94.32%, respectively. The results show that RF without any balancing effort can provide the highest detection performance on imbalanced SDN intrusion data, especially in minority classes.

One of the limitations of employing deep learning-based sampling methods such as Generative Adversarial Networks (GANs) is their training time. Deep neural networks require high optimization time due to the time-consuming backpropagation process. While these methods excel at capturing the underlying data distribution, they can overfit to training data and in some cases even memorize training instances. Simpler methods such as SMOTE are perhaps more effective in this regard as they avoid duplication and are solely based on the statistical properties of the training data.

One of the limitations in the study of intrusion detection in SDN is the lack of publicly available data from SDN. We provide benchmark results over the InSDN dataset. One future direction for this study is to extend this analysis to more datsets from SDN as more datasets become available.

Future research directions are identified as follows: An interesting direction is to employ adversarial training and/or virtual adversarial training prior to the classification of SDN intrusion data. Adversarial training consists of training classifiers (especially deep learning models) over adversarial examples that are carefully computed to be misclassified. Through adversarial training, decision boundaries are calibrated so that minority class instances would be better separated from majority cases. It would be beneficial to investigate the impact of adversarial training on classification performance, especially in the case of minority classes, where the decision boundaries are usually not as precise as those of majority classes. Other generative modeling approaches, such as Deep Convolutional Autoencoders are another interesting direction for future research as they have shown immense potential in generating realistic synthetic data based on the original data distribution.

REFERENCES

[1] P. Backs, S. Wendzel, and J. Keller, "Dynamic routing in covert channel overlays based on control protocols," in *Proc. Int. Conf. Internet Technol. Secured Trans.*, Dec. 2012, pp. 32–39.

[2] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets," *J. Big Data*, vol. 8, no. 1, pp. 1–41, Dec. 2021.

[3] G. E. Batista, R. C. Prati, and M. C. Monard, "Balancing strategies and class overlapping," in *Proc. Int. Symp. Intell. Data Anal.* Cham, Switzerland: Springer, 2005, pp. 24–35.

[4] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[5] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.

[6] K. Cabaj, L. Caviglione, W. Mazurczyk, S. Wendzel, A. Woodward, and S. Zander, "The new threats of information hiding: The road ahead," *IT Prof.*, vol. 20, no. 3, pp. 31–39, May 2018.

[7] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2010, pp. 875–886.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[9] Y.-A. Chung, H.-T. Lin, and S.-W. Yang, "Cost-aware pre-training for multiclass cost-sensitive deep learning," 2015, *arXiv:1511.09337*.

[10] D. A. Cieslak, N. V. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets," in *Proc. IEEE Int. Conf. Granular Comput.*, Sep. 2006, pp. 732–737.

[11] S. Das, S. Datta, and B. B. Chaudhuri, "Handling data irregularities in classification: Foundations, trends, and future challenges," *Pattern Recognit.*, vol. 81, pp. 674–693, Sep. 2018.

[12] M. S. Elsayed, N.-A. Le-Khac, and A. D. Jurcut, "InSDN: A novel SDN intrusion dataset," *IEEE Access*, vol. 8, pp. 165263–165284, 2020.

[13] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18–36, Feb. 2004.

[14] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[15] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[16] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456–3466, Oct. 2011.

[17] V. García, R. A. Mollineda, and J. S. Sánchez, "On the k-NN performance in a challenging scenario of imbalance and overlapping," *Pattern Anal. Appl.*, vol. 11, nos. 3–4, pp. 269–280, Sep. 2008.

[18] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 2022, pp. 1–57, Dec. 2022.

[19] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*.

[20] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," 2020, *arXiv:2008.05756*.

[21] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] H. Hindy, C. Tachtatzis, R. Atkinson, E. Bayne, and X. Bellekens, "Developing a Siamese network for intrusion detection systems," in *Proc. 1st Workshop Mach. Learn. Syst.*, Apr. 2021, pp. 120–126.

[24] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Jan. 2002.

[25] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.

[26] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018.

[27] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2. Lille, France, 2015, p. 10.

[28] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Artificial Intelligence in Medicine*. Cham, Switzerland: Springer, 2001, pp. 63–66.

[29] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, pp. 1–30, Dec. 2018.

[30] L. Liu, P. Wang, J. Lin, and L. Liu, "Intrusion detection of imbalanced network traffic based on machine learning and deep learning," *IEEE Access*, vol. 9, pp. 7550–7563, 2021.

[31] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[32] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Application of deep reinforcement learning to intrusion detection for supervised problems," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112963.

[33] K. Malialis, C. G. Panayiotou, and M. M. Polycarpou, "Data-efficient online classification with Siamese networks and active learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Mali, Jul. 2020, pp. 1–7.

[34] K. Malialis, C. G. Panayiotou, and M. M. Polycarpou, "Nonstationary data stream classification with online active learning and Siamese neural networks," *Neurocomputing*, vol. 512, pp. 235–252, Nov. 2022.

[35] W. Mazurczyk, "VoIP steganography and its detection—A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 1–21, Nov. 2013.

[36] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, nos. 2–3, pp. 427–436, Mar. 2008.

[37] I. Mirsadeghi, "Deep learning-based detection of cyberattacks in software-defined networks," in *Proc. 13th EAI Int. Conf. Digital Forensics Cyber Crime*, Boston, MA, USA, Nov. 2022, pp. 341–354.

[38] S. M. H. Mirsadeghi, H. Bahsi, and W. Inbouli, "Deep learning-based detection of cyberattacks in software-defined networks," in *Digital Forensics and Cyber Crime*, S. Goel, P. Gladyshev, A. Nikolay, G. Markowsky, and D. Johnson, Eds. Cham, Switzerland: Springer, 2023, pp. 341–354.

[39] B. Mirza, D. Haroon, B. Khan, A. Padhani, and T. Q. Syed, "Deep generative models to counter class imbalance: A model-metric mapping with proportion calibration methodology," *IEEE Access*, vol. 9, pp. 55879–55897, 2021.

[40] A. Moore, D. Zuev, and M. Crogan, "Discriminators for use in flow-based classification," Queen Mary Univ. London, London, U.K., Tech. Rep. RR-05-13, 2013.

[41] F. J. Moreno-Barea, J. M. Jerez, and L. Franco, "Improving classification accuracy using data augmentation on small data sets," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113696.
[42] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2015, pp. 1–6.
[43] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[44] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[45] M. Qiu, L. Peng, Y. Pang, B. Yang, and P. Li, "Similarity-evaluation-based evolving of flexible neural trees for imbalanced classification," *Appl. Soft Comput.*, vol. 111, Nov. 2021, Art. no. 107852.
[46] V. Raj, S. Magg, and S. Wermter, "Towards effective classification of imbalanced data with convolutional neural networks," in *Proc. IAPR Workshop Artif. Neural Netw. Pattern Recognit.* Cham, Switzerland: Springer, 2016, pp. 150–162.
[47] S. Rodda and U. S. R. Erothi, "Class imbalance problem in the network intrusion detection systems," in *Proc. Int. Conf. Electr., Electron., Optim. Techn. (ICEEOT)*, Mar. 2016, pp. 2685–2688.
[48] M. Scott and J. Plested, "GAN-SMOTE: A generative adversarial network approach to synthetic minority oversampling for one-hot encoded data," in *Proc. ICONIP*, 2019, pp. 29–35.
[49] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, "BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset," *IEEE Access*, vol. 8, pp. 29575–29585, 2020.
[50] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Jul. 2009, pp. 1–6.
[51] A. Tesfahun and D. L. Bhaskari, "Intrusion detection using random forests classifier with SMOTE and feature reduction," in *Proc. Int. Conf. Cloud Ubiquitous Comput. Emerg. Technol.*, Nov. 2013, pp. 127–132.
[52] T. Ngan, C. Haihua, J. Janet, B. Jay, and D. Junhua, "Effect of class imbalance on the performance of machine learning-based network intrusion detection," *Int. J. Performability Eng.*, vol. 17, no. 9, p. 741, 2021.
[53] S. Visa and A. Ralescu, "Learning imbalanced and overlapping classes using fuzzy sets," in *Proc. ICML*, vol. 3, 2003, pp. 97–104.
[54] P. Vuttipittayamongkol, E. Elyan, and A. Petrovski, "On the class overlap problem in imbalanced data classification," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106631.
[55] G. M. Weiss and F. Provost, "The effect of class distribution on classifier learning: An empirical study," School Arts Sci., Comput. Sci. (SAS), Rutgers Univ., New Brunswick, NJ, USA, Tech. Rep. 991031550244404646, 2001.
[56] S. J. Winham, R. R. Freimuth, and J. M. Biernacka, "A weighted random forests approach to improve predictive performance," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 6, no. 6, pp. 496–505, Dec. 2013.
[57] F. A. Md. Zaki and T. S. Chin, "FWFS: Selecting robust features towards reliable and stable traffic classifier in SDN," *IEEE Access*, vol. 7, pp. 166011–166020, 2019.
[58] H. Zhang, L. Huang, C. Q. Wu, and Z. Li, "An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset," *Comput. Netw.*, vol. 177, Aug. 2020, Art. no. 107315.
[59] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," *Comput. Intell.*, vol. 26, no. 3, pp. 232–257, 2010.

SEYED MOHAMMAD HADI MIRSADEGHI received the B.S. degree in information technology from the Institute for Advanced Studies in Basic Sciences, Iran, in 2019, and the M.Sc. degree in cybersecurity from the Tallinn University of Technology, Estonia, in 2022, where he is currently pursuing the Ph.D. degree in adversarial machine learning with the Centre for Digital Forensics and Cyber Security. His research interests include programmable networking, network security, distributed intelligence, artificial intelligence, adversarial machine learning, and deep learning.

HAYRETDIN BAHSI received the Ph.D. degree from Sabancı University, Turkey, in 2010. He is currently a Research Professor with the Centre for Digital Forensics and Cyber Security, Tallinn University of Technology, Estonia. He has more than two decades of professional and academic experience in cybersecurity. He was involved in many research and development and consultancy projects about cybersecurity as a Researcher, a Consultant, a Trainer, the Project Manager, and a Program Coordinator with the National Cyber Security Research Institute of Turkey, between 2000 and 2014. His research interests include the application of machine learning to cyber security problems, digital forensics, and cyber-physical system security.

RISTO VAARANDI received the Ph.D. degree in computer engineering from the Tallinn University of Technology, Estonia, in 2005. He is currently a Senior Researcher with the Tallinn University of Technology. From 1998 to 2018, he was affiliated with the SEB financial Group and NATO CCDCOE. His research interests include machine learning for cyber security, event correlation, data mining for event logs, and network security.

WISSEM INOUBLI received the Ph.D. degree in computer science from the Faculty of Science of Tunis, University Tunis El-Manar, Tunisia, in 2021. He is currently an Associate Professor in computer science with Artois University and a member of the Centre de Recherche en Informatique de Lens (CRIL). Previously, he was a Postdoctoral Researcher with the Lorraine Laboratory of Research in Computer Science and its Applications (LORIA), BIRD Team. Prior to that, he held a postdoctoral research position with the Data Science Group, Tallinn University. His research interests include deep learning, graph representation learning, and distributed graph mining.