Privacy-Preserving Anomaly Detection Through Sampled, Synthetic Data Generation

Fatema Rashid and Ali Miri
Department of Computer Science, Toronto Metropolitan University, Toronto, Canada

Keywords: GAN, Anomaly Detection, SMOTE, over Sampling Techniques, Under Sampling Techniques, Neural Network Classifier, Synthetic Data, TGAN.

Abstract: Anomaly detection techniques have been used successfully in various applications such as in security, financial, and medical domains. These techniques, and in particular those using advanced machine learning techniques require a high level of expertise, and the use of large volumes of data and increasing computational complexity. Outsourcing the expertise and the operational needs can provide an attractive option to many organizations. However data collected and used can include sensitive and confidential information which may require privacy protection due to legal, business or ethical considerations. We propose a novel and robust scheme that offers a flexible solution to users and organizations with varying computational and communication capabilities. Our solution would allow organizations to use semi-trusted third party cloud service providers services, while ensuring that these organizations can achieve their privacy requirement needs through the generation of synthetic data within with their computational/communication capabilities. We will demonstrate that not only does our scheme work for commonly used balanced data sets, but it is also suitable and it provides accurate results when applied to highly imbalanced data sets with extreme fluctuations in the high and low percentages of anomalies.

1 INTRODUCTION

Anomaly detection is the process of identifying *abnormal* items or events in datasets, which are different from the rest of the data, i.e. the *normal data* (Munir et al., 2019). The research in anomaly detection has been very active, due to its many potential applications in security, financial, and medical domains to name a few. Great advances have been reported in recent years, in particular by using state-of-the-art Machine Learning (ML) algorithms. However, there are still many open challenges remaining, some due to the inherent nature of anomaly detection, and others due to the setup and the application of anomaly detection in these domains. For example, many anomalies are associated with (previously) unknown behaviours, structures or distributions. Most ML-based work in the literature has focused on labeled data, i.e. supervised learning, whereas in practice it is often impractical and expensive to work with these types of datasets. It is a standard assumption that the number of anomalies is less than the normal data, although the ratio and frequency of appearance can vary greatly depending on the application and the type of anomaly. This potentially highly imbalanced data source can challenge the effectiveness of classical machine learning schemes which can show great accuracy in identifying normal data, while at the same time have poor accuracy in identifying abnormal ones. Data sources used by anomaly detection algorithm may provide rich data with many attributes. However, these attributes may be dependent/correlated, or provide irrelevant or noise-like information, i.e. the 'curse of dimensionality'. In a very recent survey by Pang *et el* (Pang et al., 2022), it has been suggested that deep learning can provide an essential role in addressing some of the common challenges listed below, and described again for completeness:

- **Low anomaly detection recall rate:** In most of data sets, anomalies are sparse and are of different types. This may result in the detection algorithm labeling normal instances as anomalies (false positives) or failing to detect some of the anomalies (false negatives). It is a challenge to reduce the false positive rate and yield high true positivity for a detection algorithm.
- **Anomaly detection in high-dimensional and/or not-independent data:** Anomalies show differ-

Rashid, F. and Miri, A.
Privacy-Preserving Anomaly Detection Through Sampled, Synthetic Data Generation.
DOI: 10.5220/0012787100003767

In *Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024)*, pages 738-747
ISBN: 978-989-758-709-2; ISSN: 2184-7711

ent behavior and characteristics when they move from low dimension to high dimension. A robust algorithm should be able to detect anomalies accurately in a high dimensional environment, where anomalies could be high order, non-linear or heterogeneous.

- **Data-efficient learning of normality/abnormality:** It is highly expensive to collect labeled data in real world scenarios. Fully supervised anomaly detection requires labeled data for training the machine learning models. To tackle this issue, it is desirable to have anomaly-detection methods which do not require labeled data.

Machine learning algorithms, and in particular deep learning technique often require complex and computationally expensive operations, as well as certain levels of expertise. Later in this paper, we will show how *Tabular Generative Adversarial Networks (TGANs)* (Xu and Veeramachaneni, 2018) can be used as part of a solution to address the above.

An extremely important design consideration for many anomaly-detection schemes is that data sources often contain sensitive and confidential information, and limited and controlled access to these types of information can be required by legal, commercial or ethical considerations. *Synthetic data generation* is an emerging area of research where artificial or synthetic data is being generated from the original data through machine learning classifiers. Synthetic data is computer generated artificial data based on user-specified parameters to ensure that the data is as close to real-world historical data as possible. Synthetic data is used nowadays for many applications such as testing and training for unprecedented scenarios, developing prototypes, etc. We would like to highlight two of its capabilities which are relevant to this paper:

- Many ML algorithms and data mining algorithms need access to huge volumes data for their operations. However, real-world data can be very expensive to collect, and in some cases this collection is restricted by law. Synthetic data can fill this need. Synthetic data can also be tuned to capture extreme and rare situational data in order to make testing more robust than with the real-world data. For example, researchers from Nvidia are teaming up with the Mayo Clinic in Minnesota and the MGH and BWH Center for Clinical Data Science in Boston to use generative adversarial networks to generate synthetic data for training neural networks. The generated synthetic data contains 3400 MRIs from the Alzheimer Disease Neuroimaging Initiative data set and 200 4D brain MRIs with tumors from the Multimodal Brain Tumor Image Segmentation Benchmark dataset (Joshi, 2022). Likewise, simulated X-rays can also be used with actual X-rays for training AI systems to recognize several health conditions (Joshi, 2022). It is important to note that the use of synthetic data allows for the implementation of advanced AI applications in areas such as healthcare and finance where the needs of the analysis must to be balanced with the need to preserve privacy.
- Machine Learning-as-a-Service (MLaaS) is a commonly used cloud service used for data processing, internal and external data sharing and big data analysis. Another capability of of synthetic data is to ensure data privacy in such settings. Synthetic data can be used to obtain desired information from data set, but made available to third parties. When organization use synthetic data as an anonymization method, a balance must be met between *efficacy* (Akcay et al., 2019) and the level of privacy protection provided. In this context, efficacy refers to the validity and the proper utilization of the data. This means that synthetic data values should provide, from an analytic point of view, the closest resemblance to real-world data values. Communication overheads, as well as computational ones listed above can pose serious challenges to use of the synthetic data as a privacy-preserving technology.

This paper will make contributions in tackling the following important questions and observations:

- How do organizations with limited or constrained computational and/or communication resources implement complex (machine learning) algorithms needed for anomaly detection?
- Anomalies typically represent a small portion of overall observed/collected data, and they may be highly variant. Complex and high-dimensional data, with possible dependency between data features can also pose additional challenges that need to be addressed. Furthermore, in practice most data collected is unlabeled, or at best partially labeled.
- Continuous monitoring and analysis required for any anomaly detection may result in exposure of confidential system and users' information to unauthorized/non-trusted parties. In these situations, the need for effective and timely detection has to be balanced with the need for privacy protection.

In the remainder of this paper, we will propose an approach that offers a flexible solution to users and organizations with varying computational and com-

munication capabilities. Our approach enables organizations to use MLaaS, while ensuring that these organizations can achieve their privacy requirements through the generation of synthetic data aligned with their computational/communication capabilities. We will demonstrate that not only does our scheme work with commonly used balanced data sets, but also it is also suitable and it provides accurate results when applied to highly imbalanced data sets with extreme fluctuations in low to high percentages of anomalies.

The rest of the paper is organized as follows. In Section II, the proposed scheme is described in detail. In Section III, the experimental results are discussed and analyzed. Related work is discussed in Section IV. Section V provides the conclusions and suggestions for avenues for future work.

2 PROPOSED SCHEME

The organization often need to make their data accessible to the third parties for analytic. This can be due to the cost or the limitation on computational resources. Communication overhead costs can also be of concern to these organizations. Anomaly detection through machine learning has been very effective in detecting anomalies, and there has been a growing list of anomaly-detection service providers in the market. In all these settings, these providers require full access to users' data to produce accurate results which could lead to system/user data privacy being compromised. In our setting, we assume that these service providers are semi-honest. That is these providers are strictly following protocols as specified by the perspective SLAs, but that they are curious about the data and as such we need privacy safeguards for data against them. Furthermore, such safeguards provide additional privacy guarantees, should the data processed or stored at these providers ever become compromised. We propose that organizations can achieve this by using sampled, synthetic data when sharing their data with service providers. We will show that our proposed scheme can accurately detect anomalies, while preserving the privacy of the underlying data.

Another technical characteristic of machine learning processes is the use of multivariate or univariate data. When the outliers are detected from the distribution of values in a single feature space, it is called *univariate anomaly detection*. For *multivariate anomaly detection*, the outliers are detected for two or more features spaces. Depending upon the data, the decision is often made to use either univariate or multivariate analysis. For our experiments, we have used both unsupervised and semi-supervised learning. The data which we used is multivariate in nature, and thus we perform multivariate analysis on this data. We have used different classifiers from different families in order to perform our experiments. We used individual detection algorithms on different datasets to obtain outlier detection:

2.1 Synthetic Data

A key component of our scheme is based on synthetic data generation. Synthetic data is generated from the original data in such a fashion that it exhibits the same underlying data distribution, characteristics and trends shown in the original data. Synthetic data has been extensively used in research purposes, and it is typically generated through statistical techniques or machine learning techniques. In this paper, we have datasets from three different domains, namely the *PIMA Indians Diabetes dataset*, the *Seismic dataset*, and the *Credit Card Fraud dataset*. The details of these datasets are presented in the Experimental Results and Analysis section.

The synthetic data properties which we are interested in are three folds, and will ensure that we can make reliable detections while respecting privacy. Firstly, the synthetic data should have the statistical properties of the original data. Secondly, it should retain the structure of the original data. The last and the most important property is that it should protect the confidentiality of the data, i.e. it must be *privacy-preserving synthetic data*.

Generative Adversarial Networks (GANs) and their variants have been among the most active subareas in deep-learning research. GANs are types of Neural Networks that are used for unsupervised learning. GANs' goal is to learn the distribution of a set of data, through the use of two opposing neural networks (Park et al., 2018). One network, the *generator* $G(x)$ creates samples that are supposed to resemble real data. The other network, the *discriminator* $D(x)$ tries to assess if a sample is real or fake based on its knowledge of the real data. After a sufficient number of iterations, the generator will produce samples that are hard to distinguish from the real ones, and hence it will learn the distribution of the data. There has been a tremendous increase in applications of GANs, including synthetic data generation. Most work so far has focused on how they have been used in image data generation. But given GANs high accuracy, and the fact that many data sets in medical, financial, and scientific fields, etc. are of a tabular nature, GANs have recently been extended to tabular data generation. In TGAN (Xu and Veeramachaneni, 2018), Xu and Veeramachaneni use a Gaussian Mixture model

and Adam optimizer in order to generate data column by column. Their model covers both discrete and continuous variables with numerical and categorical features. In this paper, we will use TGAN for our synthetic data generation for the data sets listed above.

2.2 Data Sampling Techniques

Our proposal is not only to support organizations with different computational and communication capabilities, but also with varying types of data. We are interested in scenarios such as varying ratios of abnormal to normal data and varying types of anomalies. In fact, we have selected our three datasets because of the varying degree of anomalies in them. We will achieve our goal, given these challenges through *data sampling* techniques.

Oversampling and *undersampling* in data analysis are the techniques used to adjust the class distribution of a data set. The class distribution actually represents the ratio between the different classes/categories represented. Oversampling and undersampling are opposite to each other and have different impacts on data sets when used. They should be carefully chosen depending upon the characteristics of the data set.

The most popular solution to an imbalanced data set classification problem is to change the mix of the training data sets. Techniques designed to change the class distribution in the training data sets are generally referred to as sampling methods or re-sampling methods. Oversampling techniques replicate the instances in the minority class or generate new examples from the minority class. Some of the more widely used oversampling methods include: *Random Oversampling*, *Synthetic Minority Oversampling Technique (SMOTE)* (Chawla et al., 2002), *Borderline-SMOTE* (Sun et al., 2022), *Borderline Oversampling with SVM* (Nguyen et al., 2011) and *Adaptive Synthetic Sampling (ADASYN)* (He et al., 2008).

We have used the SMOTE oversampling algorithm for our experiments due to its performance. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample as a point along that line (Mohammed et al., 2020). SMOTE generates the virtual training records by linear interpolation technique for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied to the processed data (Mohammed et al., 2020). There are many variants of the SMOTE method such as Borderline-SMOTE, Borderline Oversampling and Adaptive Synthetic Sampling that can be used in other situations.

Undersampling methods delete or select a subset of examples from the majority class to be retained in the final data set. Some of the more widely used undersampling methods are Random Undersampling, Condensed Nearest Neighbor Rule (CNN) (Batista et al., 2004), Near Miss Undersampling (Tanimoto et al., 2022), Tomek Links Undersampling (Devi et al., 2017) and the Neighborhood Cleaning Rule (NCR) (Haixiang et al., 2017). We implemented and ran our experiments on the Near Miss undersampling algorithm. This balances the class distribution of the imbalanced data sets by randomly eliminating majority class examples. When instances of two different classes are found to be very close to each other, the algorithm removes the instances of the majority class to increase the spaces between the two classes. Near Miss selects examples from the majority class that have the smallest average distance to the three closest examples from the minority class. This helps in the classification process by keeping the instances from both classes balanced and thus improves the performance of anomaly detection on imbalanced data sets.

2.3 Setup of the Proposed Scheme

Having listed the major components of our scheme, we will now discuss the setup of our proposed scheme. The typical players in our setup are the end users of an organization, the organization itself, and the cloud-service provider offering anomaly detection services to the organization. We assume that there is an explicit or implicit degree of trust between the organization and its users, but not between them and the cloud service provider which is only considered semi-honest. An example of such a setup in the financial sector is banking.

Banks often utilize third-party fraud detection services, which may require access to sensitive information, such as banking activities and users' personal information. Any compromise in confidentiality of this information could have serious consequence to both the banks and their clients.

Under our proposed scheme, organizations have different options to choose from depending on their resource capabilities. An organization with ample computation and communication capabilities can generate synthetic data using TGANs and upload the synthetic data to the anomaly detection service provider. In so doing, the detection service provider will not have access to the systems/users' sensitive information, while still being capable of performing the ana-

lytic needed to detect any abnormal data. As a baseline, we can consider an organization with no computational capability. This means that this type of organization will not be able to generate synthetic data and therefore it will have to trust the anomaly detection service provider for the privacy of their data.

An organization with limited resource capabilities does not have enough computational power to either generate entire volumes of synthetic data, yet it may not wish to upload the entire real data set due to privacy concerns. Communication overhead costs may also limit the amount of data exchanged between these types of organization and the service providers. We propose a secure and efficient solution for these types of organizations and we will evaluate the effectiveness of this solution with different types of data sets. After the initial setup and pre-training of the classifiers, this client node will generate samples from the real data. These samples are generated through oversampling and undersampling techniques. The algorithms for sample generation can produce such samples with the same characteristics and distribution as the real data. Sample generation is not a computationally intensive task and it can be easily performed with the limited capabilities of these types of organizations. These samples are smaller in volume than real time data collected by the organization. These samples are then sent to the third-party service provider, which will be responsible for performing anomaly detection and analysis. The most computationally intensive task is the synthetic data generation since it involves robust training and needs to show the classifier huge volumes of data in order to generate synthetic data mimicking the characteristics of the real data. Therefore, the limited-resource organizations have successfully outsourced the most computationally intensive tasks to the service provider, while preserving the privacy throughout the process. The process for the limited capability organizations is graphically illustrated in Figure 1. All three players and their associated roles are highlighted with different colors in Figure 1. In the remainder of this paper, we will discuss the effectiveness of our approach in terms efficacy of the cost of computation. Our experimental results support our claims that the samples generated from the real data and subsequently the synthetic data generated directly from the synthetic samples perform extremely accurately with the machine learning classifiers for anomaly detection.

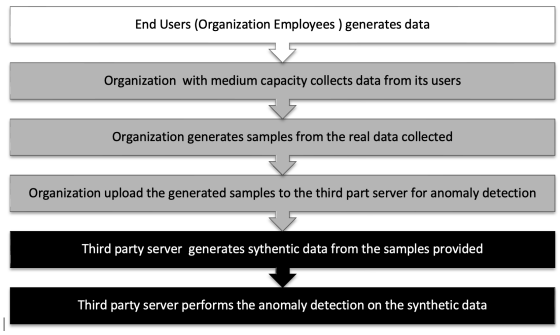

Figure 1: Flow Graph of the Proposed Scheme.

3 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we explain in detail, the datasets and classifiers we used, as well as the analysis of our experimental results. As mentioned earlier, we have used three datasets: the PIMA Indians Diabetes dataset, the Seismic Bump dataset, and the Credit Card Fraud dataset. All these datasets use a binary classification where an instance of the data is either normal or an anomaly.

PIMA DATASET WITH SMOTE SAMPLING ALGORITHM			
Classifier	ROC	Precision (0)	Precision(1)
XGBOD	0.83	0.72	0.80
iForest	0.62	0.60	0.59
kNN	0.61	0.58	0.60
HBOS	0.63	0.56	0.63
AE	0.58	0.58	0.60
Loda	0.46	0.45	0.47

PIMA DATASET WITH Near Miss SAMPLING ALGORITHM			
Classifier	ROC	Precision (0)	Precision(1)
XGBOD	0.76	0.62	0.72
iForest	0.73	0.70	0.64
kNN	0.71	0.68	0.63
HBOS	0.66	0.62	0.72
AE	0.66	0.64	0.62
Loda	0.68	0.65	0.62

Figure 2: PIMA Dataset Results.

The PIMA dataset has 9 attributes. It is composed of 35% anomalous instances and 65% normal instances. The Seismic dataset has 6.97% anomalous data and 93.3% normal instances data. It has 19 attributes in total, along with the last one indicating either an earthquake or not. It is an unbalanced data set where the positive (hazard) class is a minority class and considered to be as the outlier class and the negative class (no hazard) is considered as the normal class. The last dataset used is the Credit Card Fraud

dataset which has 0.17% anomalous data and 99.82% normal data. It has 31 attributes in total with the last one indicating either a fraud or not. The dataset contains 30 input variables. For confidentiality reasons, most of the original features have been transformed with Principal Component Analysis (PCA). This is the biggest dataset in size among the three we used. Our selected data sets include highly imbalanced data sets with extreme ranges of anomalies varying from 0.17% to 35%. The results of the performances of our proposed solutions over these three data sets are shown in Figure 2, Figure 3 and Figure 4 respectively.

SEISMIC BUMP DATASET WITH SMOTE SAMPLING ALGORITHM			
Classifier	ROC	Precision (0)	Precision(1)
XGBOD	0.71	0.52	0.86
iForest	0.54	0.53	0.54
kNN	0.74	0.51	0.72
HBOS	0.49	0.47	0.42
AE	0.43	0.52	0.42
Loda	0.64	0.54	0.60

SEISMIC BUMPS DATASET WITH NEAR MISS SAMPLING ALGORITHM			
Classifier	ROC	Precision (0)	Precision(1)
XGBOD	0.79	0.50	1.00
iForest	0.83	0.73	0.75
kNN	0.91	0.54	0.98
HBOS	0.69	0.53	0.95
AE	0.73	0.52	0.89
Loda	0.68	0.53	0.98

Figure 3: Seismic Bumps Dataset Results.

We ran our experiments on all three datasets in order to verify the performance of our proposed scheme on balanced, imbalanced and highly imbalanced data sets. All three datasets are publicly available on Kaggle.

We conducted our experiments on an Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz, 2494 Mhz, with 2 Cores, and 4 Logical Processors. We used PyOD [27], a Python library for performing anomaly detection. PyOD is a comprehensive and scalable Python toolkit for detecting outlying objects in multivariate data (Zhao et al., 2019). Since 2017, PyOD has been successfully used in various academic research and commercial products. We used 6 machine learning classifiers for performing the anomaly detection on our datasets through a synthetic data generation process. A detailed explanation of these classifiers can be found in (Zhao et al., 2019).

The first PyOD classifier used was the *XGBOD*

Credit Card Fraud DATASET WITH SMOTE SAMPLING ALGORITHM			
Classifier	ROC	Precision (0)	Precision(1)
XGBOD	0.97	0.90	0.99
iForest	0.96	0.96	0.66
kNN	0.74	0.73	0.67
HBOS	0.92	0.97	0.50
AE	0.96	0.97	0.58
Loda	0.86	0.87	0.61

Credit Card Fraud DATASET WITH Near Miss SAMPLING ALGORITHM			
Classifier	ROC	Precision (0)	Precision(1)
XGBOD	0.957	0.83	0.99
iForest	0.96	0.94	0.78
kNN	0.85	0.74	0.83
HBOS	0.952	0.95	0.70
AE	0.962	0.95	0.77
Loda	0.694	0.69	0.56

Figure 4: Credit Card Dataset Results.

classifier from PyOD. XGBOD is a semi-supervised outlier detection algorithm. It improves detection capability by creating a hybrid mix of supervised and unsupervised algorithms.

The second classifier we used is the *Auto Encoder (AE)* with Outlier Detection. AE is a type of neural networks for learning useful data representations in an unsupervised manner. AE can be used to detect outlying objects in the data by calculating the reconstruction errors.

The third classifier used was *Histograms*. Histogram-Based Outlier Score (HBOS) is an efficient unsupervised training method. It assumes feature independence and calculates the degree of outlyingness by building histograms.

The fourth classifier used was the *Isolation Forest*. The Isolation Forest separates observations by randomly selecting any specific feature and then randomly selecting a split value between the maximum and minimum values of the selected feature (Zhao et al., 2019). A recursive partitioning is used to create tree structures, with partitioning resulting in shortest paths between the root node and a terminating node indicating anomalies.

The fifth classifier we used for our implementation was the *k*-Nearest Neighbors Detector (*k*NN). For an observation, the distance to its *k*th nearest neighbor is considered to be an outlying score. It also represents a measure of density. Three *k*NN detectors are supported: (a) the *largest* which uses the distance to the *k*th neighbor as the outlier, (b) the *score mean* which uses the average of all *k* neighbors as the outlier, and

(c) the *score median* which uses the median of the distance to k neighbors as the outlier score. We used the largest kNN detector for our data sets (Zhao et al., 2019).

The sixth and the last classifier used was *Lightweight On-line Detector of Anomalies (Loda)*. Two versions of LODA are supported in PyOD: (a) *Static number of bins* which uses a static number of bins for all random cuts, and (b) *Automatic number of bins* in which every random cut uses a number of bins deemed to be optimal according to the Birge-Rozenblac method. We used the automatic number of bins method for our implementation.

As mentioned earlier in the paper, accuracy is not always the best representative of an anomaly detection algorithm performance. We have used two major performance metrics for the evaluation of the performance of our classifiers. We used area under the *Receiving Operating Characteristic curve (ROC)* and *Precision* @ rank n (P @ N) for evaluating the performance of the classifiers. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Precision rank is the fraction of relevant instances among the retrieved instances which, in our case, implies the fraction of correct anomaly detection instances among the given instances of the data.

We have presented the ROC values for when we had trained our classifier with real data (after the initialization phase where the parameters are passed from the organization to the server) and test those results against when the synthetic data generated from the samples were used. We are including the performance values of the testing phase of the models.

As we see in Figure 5 for the case of the PIMA diabetic data set, XGBOD classifier is giving us the best results when used together with the SMOTE oversampling technique. The ROC curve data point value is 0.83 with a corresponding precision of 0.80 in predicting an anomaly. This classifier gives the highest performance numbers for the PIMA dataset. We want a high number (between 0 to 1) for the ROC value. Also, the higher the precision for predicting an anomaly, the better it is. The next classifiers are HBOS, kNN and iForest respectively with successively increasing ROC values and precision to predict an anomaly. If we consider overall performance, the SMOTE sampling technique performs better than Near Miss sampling for this particular data set. As indicated earlier, 35% of data in this data set is anomalous. In short, XGBOD with SMOTE sampling is the best performer for anomaly detection in synthetic data generated from SMOTE samples for the PIMA diabetic data set.

For the Seismic Bumps dataset, the performance of the XGBOD classifier is again the best, as it was with the PIMA data set. Its ability to predict an anomaly correctly, when trained with real data and tested with synthetic data is very high. Also, the ROC values are in the range of 0.71 and 0.79 for both sampling techniques. For the Seismic bumps data set, the Near Miss undersampling technique gives slightly better performance than the oversampling technique. The next two classifiers in line are kNN and Loda for this particular data set. We would like to note again that this data set has 6.97% anomalous data. In short, the XGBOD with Near Miss sampling technique is the best performer for anomaly detection using synthetic data generated from Near Miss samples for the Seismic data set.

The next dataset is the Credit Card Fraud data set which is also highly imbalanced and has very few anomalies - as low as 0.172%. As seen in the previous data sets, XGBOD is performs best for both the undersampling and oversampling techniques. It actually has very good ability to predict an anomaly, with a probability higher than 0.50 in all cases. The ROC curve values are also above 0.80 in all cases as can be seen in Figure 5. In short, XGBOD again gives the best performance which is actually the same for both sampling techniques. The next in terms of performance are kNN and iForest classifiers, respectively.

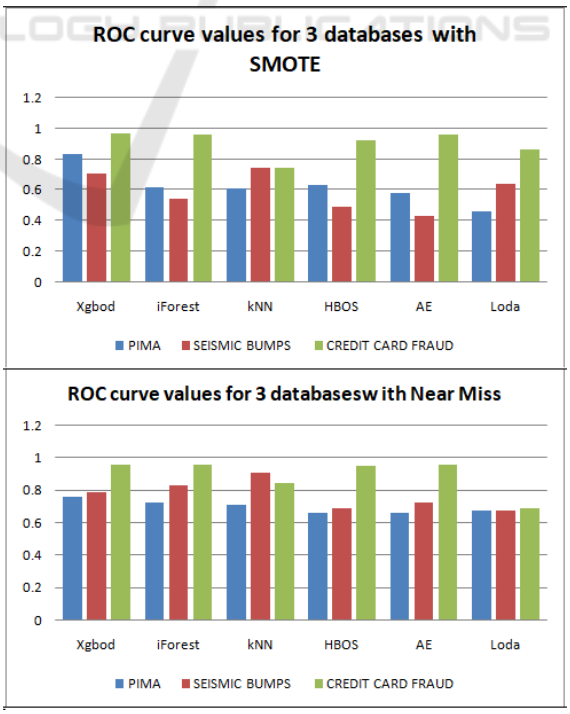

Figure 5: ROC curve values for all three data sets with Near Miss and SMOTE.

In Figure 5, it is evident that with both sampling techniques, XGBOD out performs all other classifiers. We can also conclude that the Credit Card Fraud data set exhibits the best values for the ROC curve among the other two datasets. The Credit Card dataset has the most abnormal anomaly distribution and is the most challenging dataset when it comes to anomaly detection.

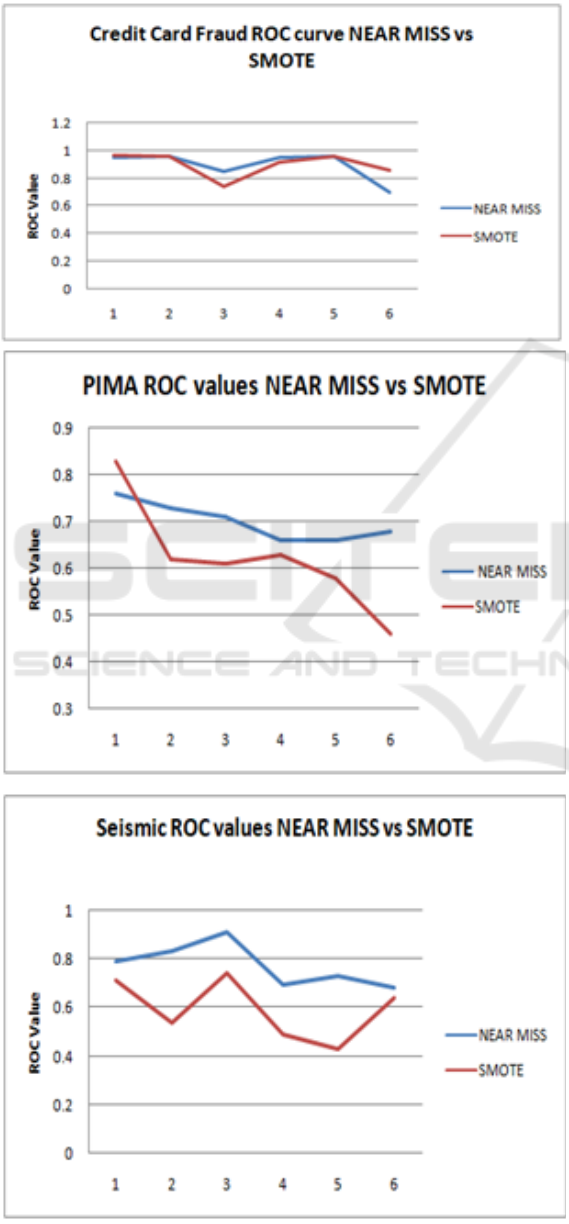

Figure 6: ROC curve values for Near Miss VS SMOTE.

In Figure 6, we plotted the performance of both the sampling techniques for each of the three datasets. SMOTE slightly outperforms Near Miss on the PIMA dataset. The difference between the performances of the sampling techniques is also small in case of the Seismic Bumps dataset. In the case of Credit Card Fraud dataset, the performance graph of both techniques overlaps each other. Therefore, we can conclude that both the undersampling and oversampling techniques perform well with all three datasets with a slight gap between the performance of the two.

Based on the above results, we can conclude that in the case of highly imbalanced and moderately imbalanced data sets, the XGBOD classifier performs the best due to being semi-supervised. The rest of the classifiers are unsupervised, and as a result they have inferior performance as a result. Our results shows that by integrating sample generation through machine learning, synthetic data generation and anomaly detection using semi-supervised classifiers, we can achieve a quite high level of accuracy and efficiency, while preserving the privacy of user data for limited resource organizations.

4 RELATED WORK

In this paper, we have proposed a novel scheme which involves the generation of samples from original imbalanced datasets and the generation of synthetic data from those samples by the anomaly detection service provider. To the best of our knowledge, there is no prior work to provide privacy-preserving anomaly detection through a combination of data sampling and synthetic data generation. Below is a brief description of some of the related work, which focus on aspects of our proposed design steps, and we have provided a comparison between their performance and functionality and that of our solution.

In (Luo et al., 2019), the authors presented the *Imbalanced Triangle Synthetic Data (ITSD)* method which attempted to provide a more general approach to generating synthetic data. They used SMOTE and its variants as their baselines, and they showed that their approach can perform better than the baselines in both precision and recall.

Using their approach, the newly formulated majority samples (SMOTE), or newly formulated minority samples (Near Miss) were added to the original data. This is in contrast to our approach of using GANs to produce a completely new synthetic data set, which provides us with greater privacy protection. Furthermore, their performance on the PIMA dataset showed an inferior F1 score.

An *ADAptive SYNthetic (ADASYN)* sampling approach was suggested in (He et al., 2008). The authors presented a comparison between their proposed

method for data generation and SMOTE with the PIMA Indian Diabetic dataset and presented a value of 0.68 for the precision of their classifier. Additional reporting on this method from PyOD library (Zhao et al., 2019) also put the values for precision and ROC in between 0.6 to 0.7 and 0.4 to 0.5, respectively. Both of these sets of numbers are inferior to what is achieved using our approach.

In (Charitou et al., 2021), a GAN-based approach called *Synthetic Data Generation GAN (SDG-GAN)* was proposed as a tool for tackling the imbalanced class problem on structured data by generating new high-quality instances. The authors tested and evaluated the SDG-GAN and compared their synthetic data generation technique with SMOTE and other methods like ADASYN(He et al., 2008). They used a number of supervised classifiers, including XGBoost. The performance of SDG-GAN when used with the PIMA and Credit Card Fraud datasets as calculated in their experiments were quite low when compared to the performance of our scheme. Also, the absence of GANs for synthetic data generation makes their scheme less accurate and secure.

In (Meng et al., 2020), the authors used the Credit Card Fraud dataset and presented the performance of the XGBoost supervised classifier for all three datasets, namely: the original dataset, the undersampled dataset, and the oversampled data using SMOTE. The values of the ROC curves were calculated as 0.97, 0.98, and 0.987, respectively, which are very close to what we have achieved in our experiments. We have used in our scheme a number of unsupervised and semi-supervised classifiers, as well as using a double layer of security by using TGANS as compared to their scheme.

In (Zaccarelli et al., 2021), a simple and efficient model based on the isolation forest algorithm for detecting amplitude anomalies on any seismic waveform segment, with no restriction on the segment record content (earthquake vs. noise) and no additional requirements than the segment metadata was presented. By considering a simple feature space composed of amplitudes of the power spectral density (PSD) of each segment evaluated at selected periods, they showed that their proposed scheme worked accurately. The evaluation results reported average precision scores of around 0.97, and maximum F1 scores above 0.9. This work did not involve any synthetic data generation, but it can present a good comparison base for the Seismic Bumps data set used for anomaly detection. We combined iForest and TGAN in our proposed scheme which achieves similar results. Our scheme thus provides more security and efficiency than their scheme in addition to its ability to provide a platform for different organizations with different capabilities to outsource their datasets to third parties for anomaly detection.

5 FUTURE WORK AND CONCLUSIONS

In this paper, we have proposed a robust privacy-preserving anomaly detection scheme, which can accommodate organizations with varying computational and communication resources. In our scheme, the organization will only need to generate samples from the real data on regular basis and send these samples to a semi-trusted party for analysis. The anomaly detection service provider will then generate the needed volume of synthetic data using the information provided by the organization, and run analytic tasks needed for detection using the generated data. Given that our approach ensures that the synthetic data mimics the same characteristics and distributions as of the real data, it can provide detection with high precision. We tested our scheme on three different data bases with a wide range of anomalies present in them. We presented a comprehensive comparison of the performance of six different classifiers with two different sampling techniques. Our experimental results using different performance metrics produced a high detection rate.

Investigating the use of Outlier Ensembles and Outlier Detector Combination Frameworks like Maximization, Average of Maximum, Maximum of Average and Median, or a combination of models to see the impact of the highly imbalanced data as compared to the individual linear models for outlier detection presented in this paper are two venues for possible future work. Further investigation is needed to ensure that TGAN or other variants of GANs are capable of distinguishing noisy data as non-anomalous. Almost all the current work in the area, including ours, has focused on (data) point anomalies. A better picture of anomalies can be found if one introduces and utilizes the concept of group or conditional anomalies. Some interesting recent work in Graph Neural Networks (GNNs) and their possible extension to anomaly detection would provide an excellent extension to this work addressing group or conditional anomalies. Most detection schemes simply classify a data point as anomalous without providing the context of how the classification decision has been made. This problem can be even more acute in the case of high dimensional data.

REFERENCES

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2019). Ganomaly: Semi-supervised anomaly detection via adversarial training. In Jawahar, C. V., Li, H., Mori, G., and Schindler, K., editors, *Computer Vision – ACCV 2018*, pages 622–637, Cham. Springer International Publishing.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29.

Charitou, C., Dragicevic, S., and d'Avila Garcez, A. (2021). Synthetic Data Generation for Fraud Detection using GANs. *arXiv e-prints*, page arXiv:2109.12546.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *SMOTE: Synthetic Minority over-Sampling Technique*, 16(1):321–357.

Devi, D., kr. Biswas, S., and Purkayastha, B. (2017). Redundancy-driven modified tomek-link based undersampling: A solution to class imbalance. *Pattern Recognition Letters*, 93:3–12. Pattern Recognition Techniques in Data Mining.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Joshi, N. (2022). BBNTimes. https://www.bbntimes.com/technology/can-synthetic-data-make-ai-better-discover-the-benefits-of-synthetic-data. Accessed on January 2024.

Luo, M., Wang, K., Cai, Z., Liu, A., Li, Y., and Cheang, C. F. (2019). Using imbalanced triangle synthetic data for machine learning anomaly detection. *Computers, Materials & Continua*, 58(1):15–26.

Meng, C., Zhou, L., and Liu, B. (2020). A case study in credit fraud detection with smote and xgboost. *Journal of Physics: Conference Series*, 1601(5):052016.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 243–248.

Munir, M., Chattha, M. A., Dengel, A., and Ahmed, S. (2019). A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 561–566.

Nguyen, H. M., Cooper, E. W., and Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21.

Pang, G., Shen, C., Cao, L., and Hengel, V. D. (2022). Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1 – 38.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083.

Sun, Y., Que, H., Cai, Q., Zhao, J., Li, J., Kong, Z., and Wang, S. (2022). Borderline smote algorithm and feature selection-based network anomalies detection strategy. *Energies 2022*, 15(13).

Tanimoto, A., Yamada, S., Takenouchi, T., Sugiyama, M., and Kashima, H. (2022). Improving imbalanced classification using near-miss instances. *Expert Systems with Applications*, 201:117130.

Xu, L. and Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. *CoRR*, abs/1811.11264.

Zaccarelli, R., Bindi, D., and Strollo, A. (2021). Anomaly detection in seismic data–metadata using simple machine-learning models. *Seismological Research Letters*, 92(4):2627–2639.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7.