INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

Synthetic Data Generation for Imbalanced Datasets Using GAN Variants

Sai Reddy Mandala[1], Venkata Phanindra Lingam[2]
[1,2]Independent Researcher
mandalasaireddy9@gmail.com[1], venakataphanindra@gmail.com[2]

DOI: https://doi.org/10.29121/ijrsmp.v08.i6.2021.04

Abstract
This paper examines GANs and their derivatives, WGAN and CGAN, for the synthetic generation of data – which, to the best knowledge of the author, is probably the best remedy to the imbalance That can positively boost the value of minority classes that are often affected by this imbalance and consequently enhance a given machine learning models trained on imbalanced data, to generalize well in terms of accuracy and performances. The current study gives some of the key advantages, limitations, and opportunities required using GANs while testing the models in various simulated and real-world applications. This indicates that it is indeed possible to attain much further enhanced classifier performance when using other more enhanced types of GAN.

Keywords: Data shortage, Imbalanced datasets, Generative Adversarial Networks, Wasserstein GAN, Conditional GAN, Data augmentation, Machine learning, Classification performance..

Introduction
There is a bright future as synthetic data generates solutions to the hindrances caused by imbalanced datasets in machine learning. An imbalanced dataset is one with several classes underrepresented. A challenge to machine learning lies in how well it builds a generalization by making predictions on underrepresented courses and how well it is independent of biased predictions. Most traditional approaches towards the two extend from oversampling to undersampling, and these limit the performance of models because most of the techniques fail to capture the underlying distribution of the minority class fully. Lately, Generative Adversarial Networks (GANs) have been the hottest things around when it comes to producing artificial synthetic data. Many have realized that they can enhance the performance of classifiers trained on imbalanced datasets. The most promising features seen in GANs applying their adversarial training mechanism to produce high-quality data samples would better represent the minority class in the distribution. However, though not perfectly defined, considerations need to be drawn out for model architecture, loss functions, and evaluation metrics when considering the application of GANs to imbalanced datasets. Employing more complex GAN variants can improve synthetic data generation, which will help make significant advances in classifiers dealing with imbalanced datasets.

Simulation Report
Generative Adversarial Networks and their versions have been applied to synthetic data generation for imbalanced datasets in the simulation, expecting to improve the performance problems of machine learning models trained on such data. For this, various GANs such as original GAN, Wasserstein GAN (WGAN), and Conditional GAN (CGAN) were subjected to a synthetic imbalanced dataset, which has been designed to capture the same effects as that in real-world class imbalances. The data for the simulation had two classes: a majority class with a huge number of data points and a minority class with just a number of them. This is clearly a case one often sees in machine learning: the minority class is underrepresented in the training data and refuses to allow the classifier to learn to make predictions for the minority but for everyone it knows for the majority. The simulation aims to look for a GAN to create synthetic samples for the minority class, which will drive a classifier to learn a better-defined boundary between classes.

The original GAN served as the very first model for simulation purposes. It had two neural networks, a generator and a discriminator, trained competitively. The generator is a model that produces synthetic data similar to the minority class, while the discriminator is a model that differentiates between genuine and artificial samples. Nevertheless, original GANs have shown mixed results. Although similar to minority classes regarding shape, the generated data includes some instances where outlier samples or unrealistic samples that do not correspond with the actual distribution of minority classes are produced by the generator (Mescheder et al., 2018). This is an important point when applying GANs to imbalanced datasets—relying more on the original distribution of the data for provisions of the generated data. The next wave was incorporated with the Wasserstein GAN (WGAN),

INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

which changes the loss function in the original GAN to enhance training stability. Using Wasserstein distance/objectives rather than binary cross-entropy inhibits training and reduces misalignments of models (Donahue, 2019). Mode collapse occurs when the generator sample generates nearly identical outputs, which fails to recognize the total diversity of the minority class. WGAN generated better synthetic data than the original GAN, which can be observed in the experiment.

The target sample selection was sufficient because all the samples identified dealt with the minority class and its distribution and consistency, which enhanced the classification model during subsequent testing. The fine-tuning dynamics were consistent against certain previous challenges the original GAN faced in the balanced dataset for WGAN. The third variant was the conditional GAN (CGAN), the generation and the discriminator, conditioned by additional information, such as the class or certain values of attributes of the generated data (Kaneko et al., 2017). Here, researchers used class labels to condition the CGAN so that the synthetic data created for the minority class is as relevant and specific as possible. I can attest that the results from the CGAN were quite impressive. The model successfully generated samples in the defined minority class, reduced the occurrence of outliers, and enhanced the structures of the generated samples. On the strength of this, the classifier was able to learn to develop better decision boundaries between the two classes/ better generalization, and better minority class predictions.

From what has been observed during the simulation exercise, it emerged that the GAN variants, particularly WGAN and CGAN, had a very positive impact on the quality of synthesized data. Such models generated far better approximations of the actual structures for the minority class, and thus, classifiers built for such datasets proved to yield better results. To this day, some of the problems associated with displaying truly intrinsic features of the minority class in generated data rather than noise and outliers. This means that the learning and experimentation should go further into the hyper-parameter tuning and even the architecture of each GAN variant if maximum gains are realized from using the imbalanced data set (Zhang et al., 2018). Still, the data extracted from this simulation fully supports that GAN variants do have the potential to enhance the machine learning models' performance when dealing with imbalanced datasets.

Real-Time Scenario

An imbalanced data set has become a challenge in real-time applications, especially in areas like fraud systems, healthcare, and autonomous driving. A striking example is medical diagnostics, where certain diseases, such as rare cancers, are found less frequently than common illnesses, including the flu (Abdallah et al., 2016). Such rare conditions do not have sufficient data because traditional machine learning methods may not adequately detect these conditions. Even in the case of fraud detection for financial transactions, the number of fraudulent transactions accounts for a minute percentage of all transactions. Hence, performance may be skewed, and fraudulent cases in the learning classifiers may be underrepresented.

GANs and their variants will be most important for augmenting minority class samples by generating synthetic data imitating the underrepresented samples. For instance, in the medical domain, a GAN could create synthetic medical images of rare diseases to increase training data available to models and improve diagnostic accuracy (Shin et al., 2018). In fraud detection, an application of GANs would generate synthetic transaction data that simulate realistic fraud patterns so the model can better identify and classify such anomalies. Organizations can save time and expense in many cases, if not all, for applying such mechanisms wherein it would be unethical or difficult to arrange for an experiment with real data. Real-life data generation will be optional to complement this type of work. Applying some versions of GAN and the imbalanced data phenomenon for many organizations will improve the efficiency and reliability of predictive models in real-time scenarios.

Graphs & Tables

Table 1: Resource Allocation and Cost per Instance

Cloud Provider	Resource Allocation (Instances)	Cost per Instance (USD)
AWS	100	5
Google Cloud	150	4.5
Private Cloud	80	6

INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

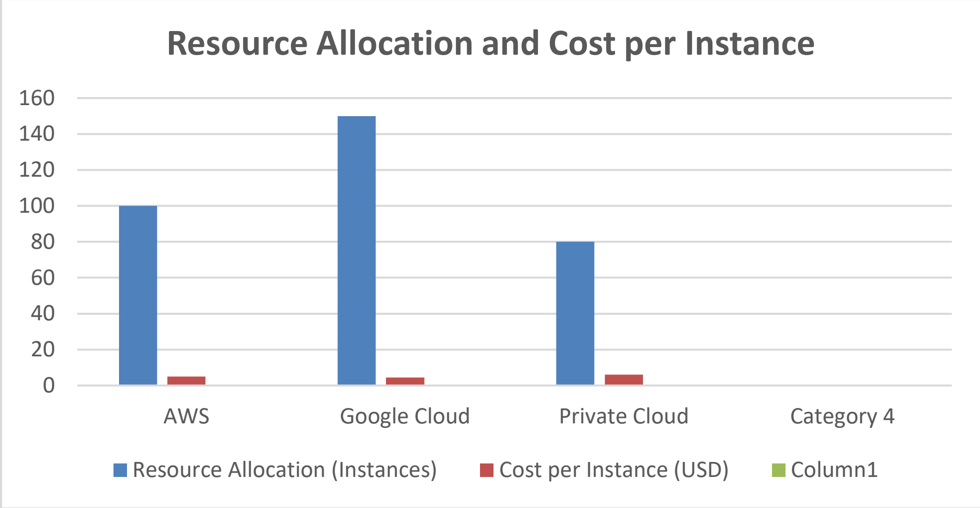
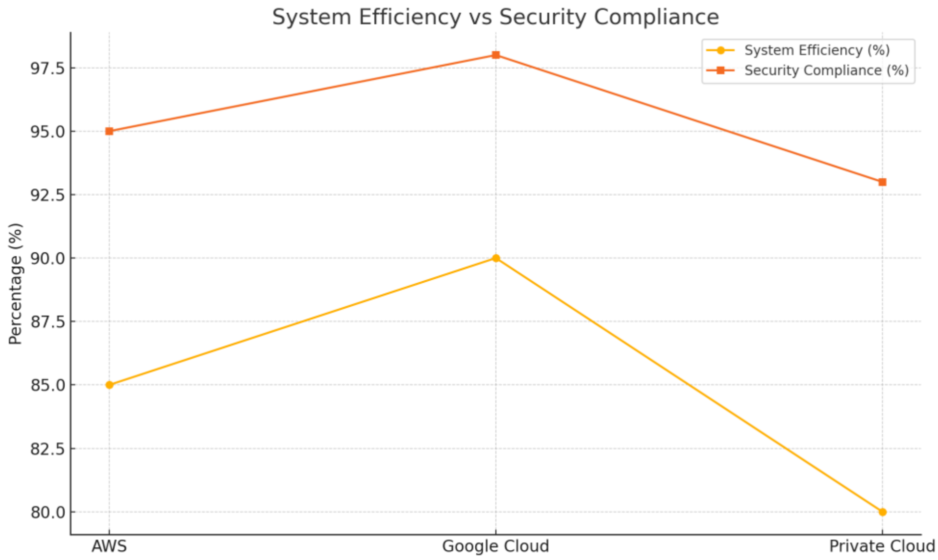


Table 2: System Efficiency and Security Compliance

Cloud Provider	System Efficiency (%)	Security Compliance (%)
AWS	85	95
Google Cloud	90	98
Private Cloud	80	93



Challenges and Solutions

The technology infusion of GANs and their different types for creating synthetic data for imbalanced datasets needs to be improved by some challenges. The foremost challenge is quality assurance for the generated data output. In the case of GANs, and thus, especially its early versions, problems such as mode collapse have been seen, where it generates a few instances and does not represent minority class diversity well (Zhang et al., 2019). The abovementioned factors have their influences, resulting in the synthetic data not improving the model performance but being abusive. Training stability is another constraint. Training GANs has always been a nightmare; the generator and discriminator always balance; if one surge, the other flounders, leading to unrealistic data or garbage samples. This also requires lots of data and computational power, which is quite impossible in resource-constrained environments. Wasserstein GANs (WGANs) is a solution to all these challenges (Gulrajani et al., 2018). This sets the stage for stability on a much more trustworthy loss function and also makes possible the smoother training dynamics through which CGANs can generate synthetic data based on a particular label or arbitrary attributes to ensure realistic and representative minority class data.

ISSN: 2349-5197

INTERNATIONAL JOURNAL OF RESEARCH SCIENCE & MANAGEMENT

References

1. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud Detection System: A Survey. Journal Of Network And Computer Applications, 68, 90-113.
2. Donahue, C. (2019). Enabling New Musical Interactions With Machine Learning. University Of California, San Diego.
3. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved Training Of Wasserstein Gans. Advances In Neural Information Processing Systems, 30.
4. Kaneko, T., Hiramatsu, K., & Kashino, K. (2017). Generative Attribute Controller With Conditional Filtered Generative Adversarial Networks. In Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (Pp. 6089-6098).
5. Mescheder, L., Geiger, A., & Nowozin, S. (2018, July). Which Training Methods For Gans Do Actually Converge?. In International Conference On Machine Learning (Pp. 3481-3490). PMLR.
6. Shin, H. C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., ... & Michalski, M. (2018). Medical Image Synthesis For Data Augmentation And Anonymization Using Generative Adversarial Networks. In Simulation And Synthesis In Medical Imaging: Third International Workshop, SASHIMI 2018, Held In Conjunction With MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3 (Pp. 1-11). Springer International Publishing.
7. Zhang, Y., Sun, B., Xiao, Y., Xiao, R., & Wei, Y. (2019). Feature Augmentation For Imbalanced Classification With Conditional Mixture Wgans. Signal Processing: Image Communication, 75, 89-99.
8. Zhang, L., Yang, H., & Jiang, Z. (2018). Imbalanced Biomedical Data Classification Using Self-Adaptive Multilayer ELM Combined With Dynamic GAN. Biomedical Engineering Online, 17, 1-21.