mathematics

MDPI

Article

Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique

Gayeong Eom [1] and Haewon Byeon [2,*]

[1] Department of Statistics, Inje University, Gimhae 50834, Republic of Korea
[2] Department of Digital Anti-Aging Healthcare (BK21), Inje University, Gimhae 50834, Republic of Korea
* Correspondence: bhwpuma@naver.com; Tel.: +82-10-7404-6969

Abstract: Classification problems due to data imbalance occur in many fields and have long been studied in the machine learning field. Many real-world datasets suffer from the issue of class imbalance, which occurs when the sizes of classes are not uniform; thus, data belonging to the minority class are likely to be misclassified. It is particularly important to overcome this issue when dealing with medical data because class imbalance inevitably arises due to incidence rates within medical datasets. This study adjusted the imbalance ratio (IR) within the National Biobank of Korea dataset “Epidemiologic data of Parkinson’s disease dementia patients” to values of 6.8 (raw data), 9, and 19 and compared four traditional oversampling methods with techniques using the conditional generative adversarial network (CGAN) and conditional tabular generative adversarial network (CTGAN). The results showed that when the classes were balanced with CGAN and CTGAN, they showed a better classification performance than the more traditional oversampling techniques based on the AUC and F1-score. We were able to expand the application scope of GAN, widely used in unstructured data, to structured data. We also offer a better solution for the imbalanced data problem and suggest future research directions.

Keywords: class imbalance; oversampling; CGAN; CTGAN; tabular data

MSC: 68T01; 68T07; 68T09

check for updates

Citation: Eom, G.; Byeon, H. Searching for Optimal Oversampling to Process Imbalanced Data: Generative Adversarial Networks and Synthetic Minority Over-Sampling Technique. *Mathematics* **2023**, *11*, 3605. https://doi.org/10.3390/math11163605

Academic Editor: Liangxiao Jiang

Received: 28 June 2023
Revised: 12 August 2023
Accepted: 19 August 2023
Published: 21 August 2023

1. Introduction

Classification issues associated with data imbalance occur in many fields and have long been studied in the machine learning field [1,2]. Many real-world datasets suffer from the class imbalance issue, which occurs because the quantities of data between classes are uneven. This issue occurs frequently in many fields, including fraud detection for credit card users [3], customer churn prediction [4], finding bad data in quality control [5], and diagnosis prediction for rare diseases [6]. In general, when machine learning techniques are used, researchers use training datasets that have similarly distributed categories with a similar sample size. When learning is conducted with imbalanced datasets, data belonging to the minority class are more likely to be misclassified than data belonging to the majority class [7]. Furthermore, even when the accuracy is high, recall or sensitivity may be low [8].

Class imbalance inevitably occurs in medical data as a function of prevalence because the amount of target data tends to be extremely small in medical contexts. For example, in cancer diagnosis, the number of patients with negative symptoms is always far greater than the number with positive symptoms. If machine learning techniques are applied without considering this bias, positive-diagnosis patients cannot be classified with high accuracy, which is problematic because the techniques will learn mainly using the negative-symptom patients who are the majority class. Furthermore, diagnosing a cancer patient as non-cancerous is much more costly than diagnosing a non-cancerous patient as having cancer. Consequently, it is critical to resolve the class imbalance problem in this case.

There are two approaches to overcoming data imbalance: the data-level approach, which manipulates data in a balanced way, and the algorithm-level approach, which responds sensitively to class imbalance [9]. Algorithm-level approaches use new variations of existing classification algorithms to solve imbalance problems [10]. A representative model for this approach is cost-sensitive learning, which defines a cost matrix to weight class misclassifications [11]. Data-level approaches adjust the sampling of the data, balancing the distribution between classes in the training data via sampling. This is a preprocessing method, and because it is implemented before the learning for classification it is independent of the classification algorithm and thus easy to apply. Therefore, data-level approaches that resolve imbalances via data sampling are more commonly studied than algorithm-level approaches that improve the learning of minority classes by adjusting the algorithm, and, thus, these oversampling techniques are often used on tabular data.

The class imbalance problem arises equally in structured and unstructured data. Among oversampling techniques based on deep learning, generative adversarial network (GAN)-based studies have recently attracted attention. GAN, a technique for generating new data by learning the distribution of the existing data, is used for unstructured data such as images, videos, and natural language processing and generally shows good performance. Although recent studies have used GAN-based oversampling for structured data, there are relatively fewer of these.

The objective of this study was to evaluate oversampling techniques for structured data, specifically tabular data with a mixture of categorical and numeric variables. Therefore, this study assessed the degree of imbalance according to the change in the imbalance ratio (IR) for clinical data. We also compared oversampling techniques which used a conditional generative adversarial network (CGAN) and a conditional tabular generative adversarial network (CTGAN) to more traditional oversampling techniques. This study aimed to find an improved solution to the imbalanced data problem as well as suggest future research directions by comparing traditional and GAN-based oversampling techniques according to the degree of imbalance after adjusting the IR to 6.8 (raw data), 9, and 19.

This paper is structured as follows: Section 1 describes the background and objectives of this research, while Section 2 presents research trends in oversampling techniques designed to deal with data imbalance. In Section 3, we describe the oversampling technique used in this paper, and in Section 4 we conduct experiments to compare the GAN-based and existing oversampling techniques using real imbalanced data. Sections 5 and 6 summarize the conclusions of the experimental results and provide directions for future research.

2. Literature Review

Data-level approaches can be classified into undersampling and oversampling techniques, depending on which class of data (by size) is controlled [12]. Studies on oversampling have mainly focused on how to generate data. Oversampling can avoid data loss by generating samples for a minority class to equalize its size with that of the majority class. However, since oversampling replicates minority-class samples, it may cause overfitting due to sample duplication, increasing the training time along with the amount of data. A diverse range of oversampling techniques exists, including random oversampling (ROS), which randomly selects and replicates samples of minority classes, the synthetic minority oversampling technique (SMOTE) [13], which generates new data using the k-nearest neighbor (k-NN) algorithm, improved SMOTE techniques (borderline-SMOTE (B-SMOTE) [14], adaptive synthetic sampling (ADASYN) [15], and the majority-weighted minority oversampling technique for imbalanced dataset learning (MWMOTE)) [16].

GAN [17] has proven its potential by generating realistic images from noise and has been utilized by studies in many fields. CGAN [18], deep convolutional GAN (DCGAN) [19], and Wasserstein GAN (WGAN) [20] have been suggested to address the problems of GAN learning and have shown excellent performance. However, since GAN evolved from image classification, there are limits to its effectiveness in generating structured data such as tables [21]. As a result, only a few studies have used GAN for the oversampling of structured data [22].

Yang et al. [23] applied CGAN to predicting drug–target interactions (DTI) and were able to balance the ratio between positive and negative samples. Oversampling methods using CGAN produce reliable samples, and these improve performance more than previous sampling methods. Quintana et al. [24] oversampled an imbalanced thermal comfort dataset using Tabular GAN (TGAN) as proposed by Xu et al. [25], who used it to oversample an imbalanced thermal comfort dataset. A particular GAN was designed to generate synthetic samples from a structured dataset, and both continuous and categorical classes were considered. Moreover, the study was able to generate both continuous and categorical data and overcome problems associated with the characteristics of tabular data by using CTGAN using the probability density for each condition, proposed by Xu et al. [26]. Wang et al. [27] applied CTGAN to traffic data to synthesize categorical samples and verify their similarity to real data, confirming that CTGAN has a higher performance and practical value than traditional oversampling and undersampling techniques. Recently, additional studies have begun to use GAN to overcome imbalance issues in structured data.

3. Materials and Methods

3.1. Imbalance Ratio (IR)

The IR can be calculated to expose the class imbalance issue using Equation (1). It can indicate how large the sample size of the majority class is compared to that of the minority class:

$$\mathrm{IR} = \frac{n^{+}}{n^{-}} \tag{1}$$

where n^{+} = number of instances in the majority class, and n^{-} = number of instances in the minority class.

When the IR is 1 or higher, a higher value indicates a greater degree of class imbalance. In particular, an IR of at least 9 indicates severely imbalanced data, with minority classes being 10% or less of the total [28].

3.2. Traditional Oversampling Techniques

3.2.1. Random Oversampling (ROS)

Random oversampling (ROS) randomly and repetitively replaces and extracts samples of a minority class until the sample sizes of the minority and majority classes become equal. Although the size of a dataset increases with the sample size of a minority class, the fact that samples of a minority class are simply replicated means it cannot be said that the amount of information increases, since the samples are duplicated. In other words, oversampling typically causes overfitting.

3.2.2. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE [13] generates samples between linearly connected structures by using the k-NN algorithm to synthesize k nearest neighbors centered on the random samples of a minority class. Since, unlike ROS, SMOTE creates samples, it has the advantage of compensating for the overfitting problem caused by duplicating the same samples. The procedure for generating synthetic samples by the SMOTE method is as follows. First, SMOTE selects samples from a random minority class for oversampling. If the number of synthetic samples to be generated is greater than the number of samples in the minority class, then all the samples in the minority class are selected. If it is less, then a subset of all the samples in the majority class is randomly selected. Second, k nearest neighbors are selected around the minority-class random samples that have a linear relationship with the minority-class samples based on the first step; these are then multiplied by the weight, and a synthetic sample is created at the location of the multiplied value. A mathematical representation of this is given in Equation (2):

$$x_{smote} = x_i + (\hat{x}_i - x_i) \times \delta, \quad \delta \in [0,\ 1], \quad i = 1, 2, \cdots, k \tag{2}$$

where x_i is a sample belonging to a minority class and $\hat{x}_i$ is a random neighbor among k-NNs for x_i. The process works by identifying the k nearest neighbors near x_i, calculating the differences between x_i and these neighbors, and multiplying by a value between 0 and 1 to create a synthetic sample x_{smote} to supplement the original samples. This is repeated until the size of the minority class becomes equal to that of the majority class.

3.2.3. Borderline-SMOTE (B-SMOTE)

B-SMOTE [14] is an expansion of SMOTE. Whereas SMOTE generates a composite sample of the minority class without considering the location of neighboring samples, B-SMOTE defines the region where the two classes overlap as the boundary and applies the SMOTE technique to the minority-class samples on the boundary to generate a composite sample.

The B-SMOTE procedure is as follows. First, for each individual sample belonging to a minority class, the k closest observations are found, regardless of the class. Second, if S_{maj} is the sample size of the majority class, it is classified as a "Danger" group if $\frac{k}{2} \leq S_{maj} < k$, as a "Safe" group if $0 \leq S_{maj} < \frac{k}{2}$, and as a "Noise" group if $S_{maj} = k$. Third, this method generates new samples only for minority-class samples belonging to the "Danger" group.

3.2.4. Adaptive Synthetic Sampling (ADASYN)

ADASYN [15] is an advanced form of SMOTE that calculates the density distribution for each sample of a minority class and determines the number of samples to be generated accordingly. ADASYN creates synthetic samples as follows. First, it finds K nearest neighbors for sample x_i belonging to a minority class S_{min} and denotes the number of samples belonging to the minority class as Δ_i. Then, it calculates an r_i, density distribution, which can be expressed as Equation (3), while in Equation (4), $\hat{r}_i$ refers to the normalized r_i:

$$r_i = \frac{\Delta_i}{K}, \quad i = 1, 2, \ldots, S_{min} \tag{3}$$

$$\hat{r}_i = r_i / \sum_{i=1}^{S_{min}} r_i \tag{4}$$

In Equation (5), G calculates the number of samples to be generated for S_{min} and β is used to balance the samples between the two classes:

$$G = \left(S_{maj} - S_{min}\right) \times \beta, \beta \in [0, 1] \tag{5}$$

Next, ADASYN determines the number of synthetic samples (g_i) that need to be generated for samples x_i belonging to S_{min} and generates these samples by repeating the process g_i times for each x_i. Equation (6) expresses this as the formula:

$$g_i = r_i \times G. \tag{6}$$

3.3. GAN-Based Oversampling Technique

3.3.1. Generative Adversarial Network (GAN)

GAN [17] is a deep learning-based unsupervised learning model that generates fake data resembling real data by pitting one neural network (generator, G) against the other (discriminator, D). G is trained with the goal of producing fake data that resemble real data, while D is trained to determine that the data created by G is indeed fake. In other words, G and D learn in an adversarial way.

Figure 1 depicts the structure of GAN and describes how G and D learn. First, when G receives a random noise vector as input, data are generated. When the generated and actual data are provided to D, it determines whether they are real or fake. G and D compete for and learn from this result. Put differently, the goal of G is to maximize the probability that D determines the generated data as real, while the goal of D is to maximize the probability of discriminating generated data as fake. Equation (7) shows the GAN's objective function for this learning process:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data}(x)}[logD(x)]+E_{z\sim p_z(z)}[log(1-D(G(z)))]. \tag{7}$$

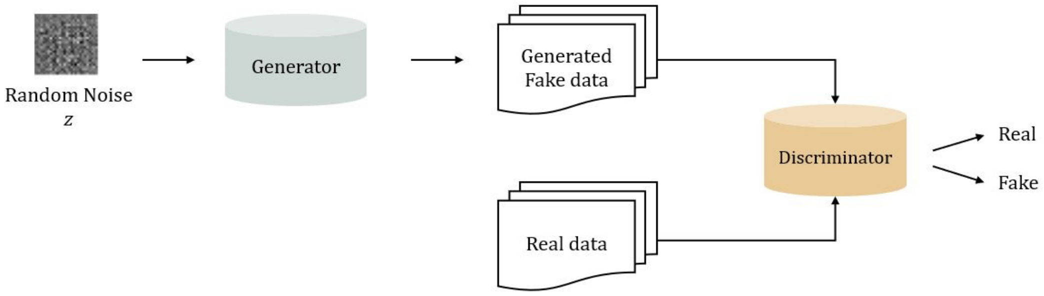

Figure 1. Structure of GAN.

In the equation, $p_{data}(x)$ and $p_z(z)$ refer to the real and fake data, respectively. D receives data x as a real-data input value and outputs the probability of being real data $(D(x))$. G takes a random noise vector z as an input value and generates fake data $(G(z))$. Since the goal of D is to distinguish effectively between generated fake data and real data, the GAN must learn so that $D(x)$ is 1 and $D(G(z))$ is 0. At the same time, since the goal of G is to deceive D, it should learn to make $D(G(z))$ equal to 1. In other words, the objective function of the equation aims at maximization from the perspective of D and minimization from the viewpoint of G.

3.3.2. Conditional GAN (CGAN)

CGAN [18] is designed to improve the unstable learning of GAN. Although the basic learning method is the same, CGAN can impact the data generation process directly and learn characteristics as well as distribution by adding a feature y, which indicates a specific condition, to G and D.

Figure 2 shows the process of generating data. It enters y, the feature desired by the user, along with a random noise vector z; the information for the labeled class is input to D. The objective function for learning CGAN is the same as for GAN, but y is conditionally added as expressed as Equation (8):

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data}(x)}[logD(x|y)]+E_{z\sim p_z(z)}[log(1-D(G(z|y)))]. \tag{8}$$

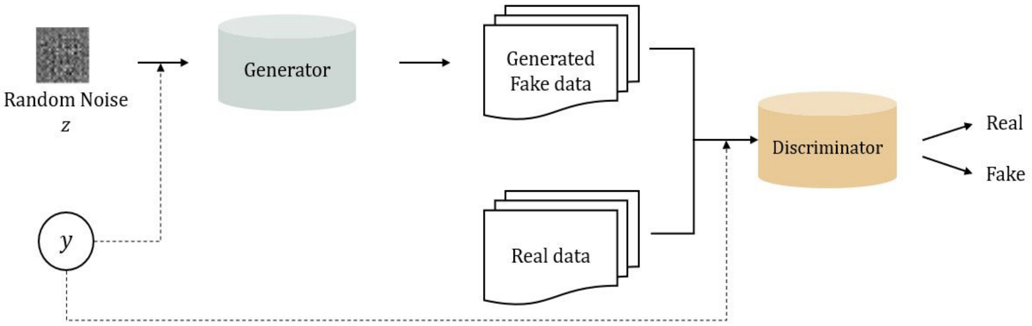

Figure 2. Structure of CGAN.

3.3.3. Conditional Tabular GAN (CTGAN)

Conventional GAN algorithms have shown strong performance in the process of learning original images and generating and predicting synthetic images for each condition [18,21]. However, it is difficult to apply them to structured data, constituting a shortfall [18,21], as they suffer from problems such as various tabular data types, data distributions not following the Gaussian distribution, multi-modal data types, sparse matrices generated by one-hot encoding, and categorical variables with a high degree of imbalance [29].

As a result, CTGAN [26], a generative model designed to use GAN functions for structured data, was proposed. CTGAN is a model that combines the conditional-GAN [18] and the tabular-GAN algorithms [25]. A common problem with GANs is that they do not learn sparse categories well if certain categories are imbalanced. Therefore, CGANs allow for the adding of conditions to the constructor to ensure that sparse categories are included in the learning process.

CTGAN proposes mode-specific normalization and training-by-sampling to solve the problems caused by GANs. Mode-specific normalization, a component of CTGAN, learns while considering multimodal and non-Gaussian distribution problems by normalizing numerical data using the variational Gaussian mixture. In the learning process, the normalized values of each numerical variable are used as the input, rather than the values of the original data. After learning is completed, the data created through G are converted to the scale of the original data.

Training-by-sampling is a method for uniformly sampling the state vector and the training data. To make the conditional distributions of the constructor representation and the actual data equal, the difference between the two distributions must be accurately estimated from the identifier (Critic). The specific procedure is shown in Figure 3. The procedure for training by sampling is as follows. First, select one of the categorical columns with equal probability and take a logarithmic function of the frequency of each category to create a probability distribution over the frequency of occurrence of each value. Second, generate a state vector according to the selected column and class and randomly sample the data for training so that the generator generates a representation through the state vector and the latent variable. Third, by putting the actual data and the reproduced data into an identifier and calculating the distance (score) between the two conditional distributions, the generator learns to generate data that satisfies the conditions.

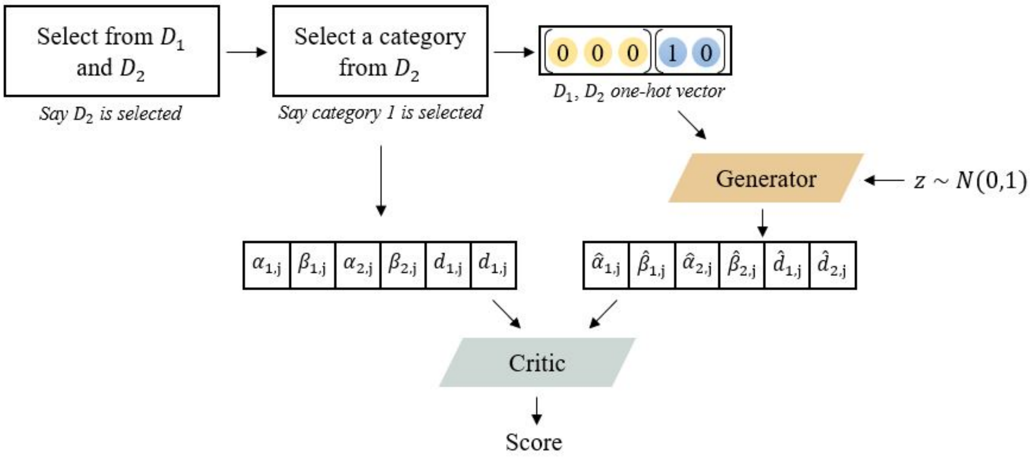
Select from D_1 and D_2
Say D_2 is selected
Select a category from D_2
Say category 1 is selected
0 0 0 1 0
D_1, D_2 one-hot vector
Generator
$z \sim N(0,1)$
$\alpha_{1,j}$ $\beta_{1,j}$ $\alpha_{2,j}$ $\beta_{2,j}$ $d_{1,j}$ $d_{1,j}$
$\hat{\alpha}_{1,j}$ $\hat{\beta}_{1,j}$ $\hat{\alpha}_{2,j}$ $\hat{\beta}_{2,j}$ $\hat{d}_{1,j}$ $\hat{d}_{2,j}$
Critic
Score

Figure 3. Training-by-sampling of CTGAN.

3.4. Data

The data source for this study is the "Epidemiologic data of Parkinson's disease dementia patients" from the National Biobank of Korea under the Korea Disease Control and Prevention Agency. Data were collected from 14 tertiary medical institutions (university hospitals) nationwide from January 2015 to December 2015 under the supervision of the Korea Centers for Disease Control and Prevention. A health survey was conducted using computer-assisted personal interviews (CAPI). We obtained the approval of the Korea Disease Control and Prevention Agency's Research Ethics Review Committee (No. KBN-2019-005) and the National Biobank Korea's Lotting-out Committee (No. KBN-2019-1327) before abstracting and analyzing the data.

The data contain information on Alzheimer's disease and Parkinson's disease patients. The data classify Parkinson's disease patients into dementia, mild cognitive impairment, and normal cognitive function. The explanatory variables consist of 54 variables such as basic information, environmental factors, disease history, Alzheimer's/Parkinson's disease basic information, and clinical scale. Fourteen continuous variables and seven categorical variables were selected as the final explanatory variables based on using feature importance. Missing values for each item were replaced by mean imputation.

The dependent variable, "patient classification", reclassified Parkinson's disease patients into two classes after excluding Alzheimer's patients: 0 means Parkinson's disease patients with normal cognitive function (51 patients) and 1 means Parkinson's disease patients with dementia or mild cognitive impairment (125 and 223 patients, respectively). Out of the 399 Parkinson's disease patients, 51 had normal cognitive function, accounting for 12.78% of the total data. This yielded an IR value of 6.8, indicating the presence of an imbalance. Table 1 shows the numbers and ratios by category.

Table 1. Description of the dependent variable.

	Normal Cognitive Function	**Dementia and Mild Cognitive Impairment**	**Total**
Sample	51	348	399
Ratio	12.78%	87.22%	100%

3.5. Experimental Design

First, IR values were adjusted to 6.8 (raw data), 9, and 19 for comparing oversampling techniques according to the imbalance ratio. In case of insufficient data in the majority class, data were created with CTGAN specialized for structured data and added to the original data to prevent data loss. Minority classes were randomly extracted from the original data as needed. The numbers of samples according to the IR value are shown in Table 2.

Table 2. Numbers of samples by IR value.

	IR = 6.8 (Raw Data)	**IR = 9**	**IR = 19**
Normal cognitive function	51	40	19
Dementia and mild cognitive impairment	348	359	380

This study used ROS, SMOTE, B-SMTOE, and ADASYN techniques, comparing them with oversampling techniques using GAN and CTGAN. The imblearn package was used for this purpose. Moreover, k = 5 was used for k-NN-based SMOTE, B-SMOTE, and ADASYN. Sampling was adjusted to make the ratio of normal cognitive function (0) and dementia and mild cognitive impairment (1) equal to 1:1. Numbers could vary slightly because, unlike other oversampling techniques, ADASYN oversampled by automatically adjusting the number as needed in the package.

When learning CGAN, both G and D consisted of three hidden layers with the epoch set to 1000. In addition, Leaky ReLU and Adam were used as activation functions and optimizers, respectively; Adam's learning rate, β1, and β2 were set to 0.0002, 0.5, and 0.9, respectively. CTGAN is usable in the Synthetic Data Vault (SDV) [30], and the experiment was conducted by setting the epoch to 100. The amount of data after applying each oversampling to the dataset is shown in Table 3. The support vector machine (SVM) [31], logistic regression (LR) [32], random forest (RF) [33], and multi-layer perceptron (MLP) [34] were used as classification models.

Table 3. Amounts of data after applying each oversampling to the dataset.

Technique	**Total Sample (Normal vs. Cognitive Impairment)**		
	IR = 6.8	**IR = 9**	**IR = 19**
CTGAN CGAN ROS SMOTE B-SMOTE	696 (348:348)	718 (359:359)	760 (380:380)
ADASYN	702 (354:348)	722 (363:359)	

3.6. Performance Evaluation Methods and Indicators

The entire dataset was divided, with 80% used as training data and the remaining 20% as validation data. This study conducted a 10-fold cross-validation to circumvent the problem of greatly varying model performance by chance and determined the final performance based on the mean performance of the ten models.

The F1-score and area under the curve (AUC), widely used in class imbalance studies, were used as indicators for performance evaluation [35,36]. In the confusion matrix, TP and TN indicate true positive (predict positive for positive) and true negative (predict negative for negative), respectively. FP and FN stand for false positive and false negative, respectively. The F1-score is the harmonic mean of precision and recall; the closer it is to 1, the better the classification performance of the minority class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{TPR}(\text{True Positive Rate}) \tag{10}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

The "receiver operating characteristic (ROC) curve" means a curve presenting the classification prediction result of the model with the TPR (true positive rate, recall) on the vertical axis and FPR (false positive rate, 1-specificity) on the horizontal axis. AUC is the area under the ROC curve; the closer it is to 1, the better the performance; it is calculated as the mean of TPR and TNR.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \text{TNR}(\text{True Negative Rate}) \tag{12}$$

$$\text{AUC} = \frac{\text{TPR} + \text{TNR}}{2} \tag{13}$$

4. Results

This study presents the result of comparing classification performances when applying each oversampling technique after adjusting the IR of the experimental dataset (IR = 6.8) to 9 and 19. Values showing the best and lowest performance are bold and underlined, respectively.

Table 4 shows the AUC scores for the classification results. The GAN-based oversampling techniques showed a higher performance than the traditional oversampling techniques in all areas. CTGAN showed a strong performance, especially in the SVM and

LR classification models. CGAN produced high AUC scores in the MLP and RF classification models. Although the ROS technique exhibited the poorest performance among the traditional oversampling techniques, it did not do so in the LR because CGAN, which generally performed well, rapidly fell away in performance with the LR classification model. Moreover, although CTGAN showed higher AUC scores than conventional oversampling techniques in SVM, LR, and RF, it was confirmed that its performance decreased under classification by MLP.

Table 4. Comparison of performance by oversampling technique (AUC).

Mean of AUC Scores						
IR	**ROS**	**SMOTE**	**ADASYN**	**B-SMOTE**	**CGAN**	**CTGAN**
Classification model: SVM						
6.8	0.8038	0.7942	0.7752	0.8226	0.8248	**0.8329**
9	0.7959	0.8393	0.8176	0.8285	0.8291	**0.8488**
19	0.7735	0.8430	0.8169	0.8415	0.8540	**0.8609**
Classification model: LR						
6.8	0.7851	0.8061	0.7882	0.8110	0.7486	**0.8165**
9	0.8100	0.8241	0.8093	0.8274	0.7935	**0.8342**
19	0.8130	0.8202	0.8081	0.8348	0.8109	**0.8452**
Classification model: RF						
6.8	0.8708	0.8998	0.8971	0.9051	**0.9484**	0.9200
9	0.8876	0.9086	0.9024	0.9072	**0.9550**	0.9340
19	0.8889	0.9222	0.9067	0.9332	**0.9750**	0.9470
Classification model: MLP						
6.8	0.8339	0.8717	0.8387	0.8581	**0.8896**	0.8420
9	0.8492	0.8567	0.8613	0.8800	**0.9177**	0.8667
19	0.8706	0.9075	0.8910	0.9290	**0.9482**	0.8970

In other words, the combinations CTGAN + SVM, CTGAN + LR, CGAN + RF, and CGAN + MLP showed the highest performance, while the ROS method showed the lowest performance in most classification models. Moreover, despite the increase in the degree of imbalance, no technique showed greatly decreased performance. Rather, the overall classification performance increased slightly.

Table 5 shows the F1-score values for the classification results. As with the AUC score, the GAN-based oversampling technique showed a better performance than the traditional oversampling technique. Effective performance can be seen in the combinations CTGAN + SVM, CTGAN + LR, CGAN + RF, and CGAN + MLP, while the ROS method showed the lowest overall performance. All methods showed stable performance, even for higher IR values.

Figures 4 and 5 display the AUC and F1-score values for six oversampling techniques by classification model. Ranking the techniques revealed differences, even though the classification performances for the techniques appeared similar when only the best and lowest AUC and F1-score performances were examined. This is because the two measures indicate different things. Even considering these elements, the results of this study confirmed that CGAN and CTGAN showed better AUC and F1-score results than the existing oversampling techniques.

Table 5. Comparison of performance by oversampling technique (F1-score).

Mean of F1-Scores						
IR	**ROS**	**SMOTE**	**ADASYN**	**B-SMOTE**	**CGAN**	**CTGAN**
Classification model: SVM						
6.8	0.7347	0.7264	0.7083	0.7447	0.7673	**0.7873**
9	0.7493	0.7948	0.7688	0.7724	0.7660	**0.8175**
19	0.7016	0.7704	0.7614	0.8122	0.8212	**0.8302**
Classification model: LR						
6.8	0.7663	0.7758	0.7619	0.7749	0.7178	**0.7856**
9	0.7973	0.7903	0.7768	0.7914	0.7689	**0.8128**
19	0.7797	0.7885	0.7869	0.8086	0.7947	**0.8214**
Classification model: RF						
6.8	0.8497	0.8813	0.8861	0.8846	**0.9362**	0.9050
9	0.8604	0.8945	0.8858	0.8812	**0.9465**	0.9190
19	0.8720	0.9110	0.8914	0.9236	**0.9724**	0.9369
Classification model: MLP						
6.8	0.8113	0.8555	0.8224	0.8401	**0.8710**	0.8289
9	0.8221	0.8359	0.8472	0.8488	**0.9064**	0.8461
19	0.8366	0.8862	0.8815	0.9121	**0.9408**	0.8833

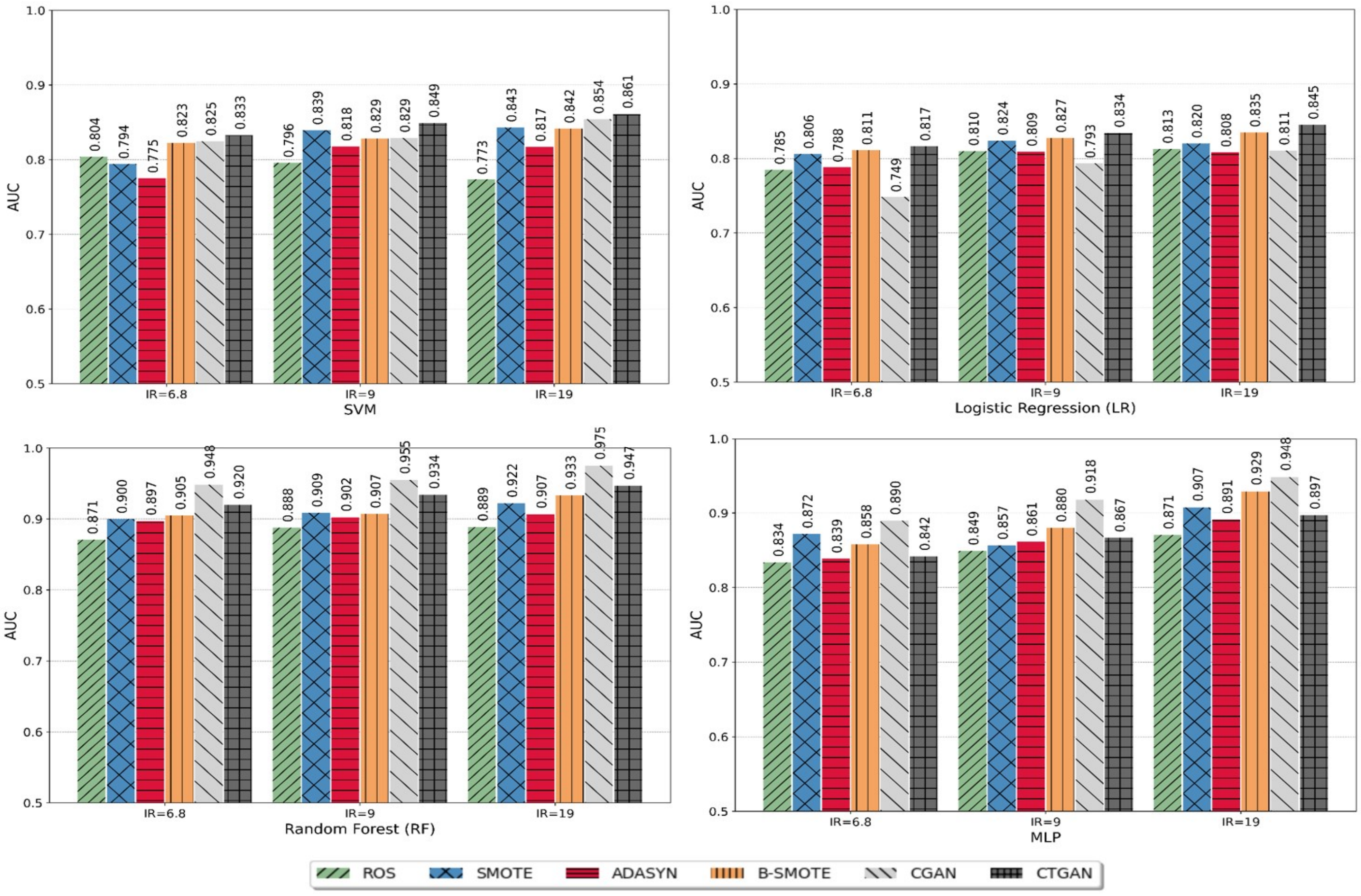

Figure 4. Bar chart of the six oversampling techniques (AUC).

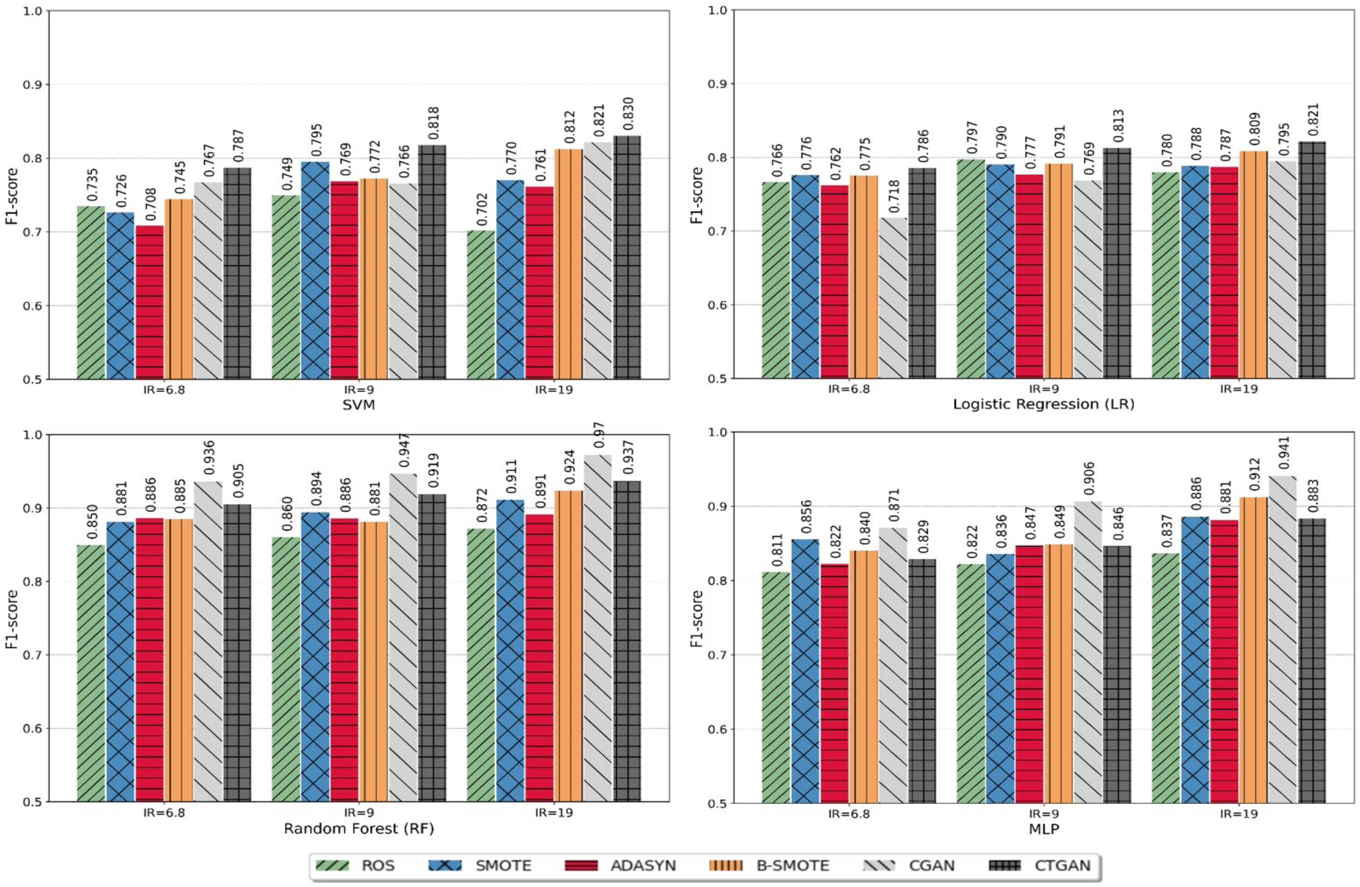

Figure 5. Bar chart of the six oversampling techniques (F1-score).

5. Discussion

Most medical datasets have class imbalance issues due to low incidence rates. It is very important to overcome this issue because the misclassification of a minority class can decrease sensitivity among classification performance components. Therefore, this study adjusted the imbalance ratio (IR) to 6.8 (raw data), 9, and 19 using actual epidemiological data on Parkinson's disease dementia patients. The study applied oversampling techniques using CGAN and CTGAN as well as more traditional oversampling techniques (ROS, SMOTE, ADASYN, and B-SMOTE); it aimed to solve the imbalance problem by comparing the performance of each technique through classification models (SVM, LR, RF, and MLP).

This study classified the levels of cognitive impairment associated with Parkinson's disease by applying oversampling techniques to three datasets with three different IR values and found that GAN-based oversampling techniques showed better AUC and F1-score values than traditional techniques. Nugraha et al. [37] used insurance fraud imbalance data and proposed CTGAN as an oversampling method, showing that over the application of 17 classification models, CTGAN presented a better performance (AUC, F1-score, precision, etc.) than ROS, SMOTE, and ADASYN. A study using imbalanced CVD clinical data by García-Vicente et al. [38] also found that the combination of CTGAN and the classification model LASSO showed strong potential for generating categorical data. Many previous studies [39–41] also showed that the CGAN-based oversampling technique achieved a higher performance than more traditional techniques over various classification models for datasets with complex structures because it was effective at generating data for the minority class, whereas the traditional minority oversampling techniques added data randomly rather than based on the actual data distribution, a significant limitation of these techniques [42]. Moreover, SMOTE-based oversampling techniques are ineffective at reproducing high-dimensional data, being more useful for low-dimensional data, i.e., another shortcoming. In contrast, previous studies [43] reported that GAN could overcome the disadvantages of existing oversampling techniques because it generated data according to the distribution of actual data and was effective even for high-dimensional data. The results of this study also demonstrated that GAN treated imbalanced data better than more traditional minority oversampling techniques such as SMOTE in high-dimensional data. More recently, Sharma et al. [44] developed a SMOTified-GAN algorithm, a data augmentation technique based on variations of GAN designed to overcome the class imbalance classification problem. However, future studies would be useful to evaluate the effectiveness of GAN on various imbalanced datasets.

The significance of this study lies in its confirmation that the combinations CTGAN + SVM, CTGAN + LR, CGAN + RF, and CGAN + MLP showed better performance, proving that GAN-based oversampling contributed to improving classification accuracy in clinical data by comparing the classification performance of various oversampling techniques. The study also demonstrated that CTGAN oversampling could generate high-quality synthetic data without adjusting any hyperparameter. As a result, it will be possible to expand the application scope of GAN, which has been widely used for unstructured data such as images and videos.

This study had several limitations. First, since only one dataset was used, the study could not compare the performance of oversampling techniques according to dataset size or the ratio of categorical to continuous variables. Second, the optimal number of epochs in the process of learning CTGAN could not be determined. It was, therefore, necessary to learn many times, thus, the best performance might not be identifiable due to the optimal number of learning times not being known. Third, there were many missing values (e.g., answered as "don't know") due to the nature of medical data. Moreover, there was little change in performance over variations in the IR because the sample size was small; so, there was not much difference in the sample size of the minority class according to the degree of imbalance due to the use of actual Parkinson's disease patient data. Future studies should aim to identify oversampling techniques more accurately by applying oversampling to multiple datasets and checking the difference in classification performance while taking this into account.

6. Conclusions

This study confirmed the effectiveness of CTGAN and CGAN oversampling techniques by applying six oversampling techniques to imbalanced data and comparing their performance. Data imbalance is a critical problem because it occurs in many fields, including the medical field featured in this study. It should be possible to apply the superior performance of GAN-based oversampling to imbalance issues based on the study's results. Future studies need to identify the optimal oversampling technique by comparing its performance to the performance of other types of techniques in addition to more traditional oversampling techniques.

Author Contributions: Conceptualization, G.E. and H.B.; software, G.E.; methodology, G.E. and H.B.; validation, G.E. and H.B.; investigation, G.E.; writing—original draft preparation, G.E.; formal analysis, H.B.; writing—review and editing, G.E. and H.B.; visualization, G.E. and H.B.; supervision, H.B.; project administration, H.B; funding acquisition, H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, No. RS-2023-00237287 and NRF-2021S1A5A8062526 and local government–university cooperation-based regional innovation projects (2021RIS-003).

Institutional Review Board Statement: The study was carried out in accordance with the Helsinki Declaration and was approved by the Korea Workers' Compensation and Welfare Service's Institutional Review Board (or Ethics Committee) (protocol code 0439001, date of approval 31 January 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data source for this study was the 'Epidemiologic data of Parkinson's disease dementia patients' of the National Biobank of Korea (https://nih.go.kr/biobank/cmm/main/mainPage.do; accessed on 15 June 2023.) under the Korea Disease Control and Prevention Agency. In order to use this data, it is necessary to obtain approval from the Korea Centers for Disease Control and Prevention Research Ethics Review Committee and the Lotting-out Committee of the National Biobank of Korea.

Conflicts of Interest: The authors declare that there are no conflict of interest regarding the publication of this article.

References

1. Chen, Z.; Duan, J.; Kang, L.; Qiu, G. Class-Imbalanced Deep Learning via a Class-Balanced Ensemble. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 5626–5640. [CrossRef] [PubMed]
2. Xie, Y.; Qiu, M.; Zhang, H.; Peng, L.; Chen, Z. Gaussian Distribution Based Oversampling for Imbalanced Data Classification. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 667–679. [CrossRef]
3. Phua, C.; Alahakoon, D.; Lee, V. Minority Report in Fraud Detection: Classification of Skewed Data. *ACM SIGKDD Explor. Newslett.* **2004**, *6*, 50–59. [CrossRef]
4. Hung, S.Y.; Yen, D.C.; Wang, H.Y. Applying Data Mining to Telecom Churn Management. *Expert Syst. Appl.* **2006**, *31*, 515–524. [CrossRef]
5. Kim, A.; Oh, K.; Jung, J.Y.; Kim, B. Imbalanced Classification of Manufacturing Quality Conditions Using Cost-Sensitive Decision Tree Ensembles. *Int. J. Comput. Integr. Manuf.* **2018**, *31*, 701–717. [CrossRef]
6. Mazurowski, M.A.; Habas, P.A.; Zurada, J.M.; Lo, J.Y.; Baker, J.A.; Tourassi, G.D. Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance. *Neural Netw.* **2008**, *21*, 427–436. [CrossRef]
7. Ertekin, S.; Huang, J.; Bottou, L.; Giles, L. Learning on the Border: Active Learning in Imbalanced Data Classification. In Proceedings of the ACM International Conference on Information and Knowledge Management, Lisbon, Portugal, 6–10 November 2007; pp. 127–136. [CrossRef]
8. Lee, H.; Hong, S.; Bang, J.; Kim, H. Study of Optimization Techniques to Apply Federated Learning on Class Imbalance Problems. *J. Korea Inst. Inf. Technol.* **2021**, *19*, 43–54. [CrossRef]
9. Lee, K.; Lim, J.; Bok, K.; Yoo, J. Handling Method of Imbalance Data for Machine Learning: Focused on Sampling. *J. Korea Contents Assoc.* **2019**, *19*, 567–577. [CrossRef]
10. Wen, G.; Li, X.; Zhu, Y.; Chen, L.; Luo, Q.; Tan, M. One-Step Spectral Rotation Clustering for Imbalanced High-Dimensional Data. *Inf. Process. Manag.* **2021**, *58*, 102388. [CrossRef]
11. Elkan, C. The Foundations of Cost-Sensitive Learning. In *International Joint Conference on Artificial Intelligence*; Lawrence Erlbaum Associates Ltd.: Mahwah, NJ, USA, 2001; Volume 17, pp. 973–978.
12. Van Hulse, J.; Khoshgoftaar, T.M.; Napolitano, A. Experimental Perspectives on Learning from Imbalanced Data. In Proceedings of the 24th International Conference on Machine Learning, Corvalis, OR, USA, 20–24 June 2007; pp. 935–942. [CrossRef]
13. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
14. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *International Conference on Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887. [CrossRef]
15. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008; pp. 1322–1328. [CrossRef]
16. Barua, S.; Islam, M.M.; Yao, X.; Murase, K. MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 405–425. [CrossRef]
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
18. Mirza, M.; Osindero, S. Conditional Generative Adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
19. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
20. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
21. Hwang, K. R&D Accountability and Dilemma within the Korean Science and Technology Context. *Korean Public Adm. Rev.* **2016**, *50*, 189–213.
22. Engelmann, J.; Lessmann, S. Conditional Wasserstein GAN-Based Oversampling of Tabular Data for Imbalanced Learning. *Expert Syst. Appl.* **2021**, *174*, 114582. [CrossRef]
23. Yang, K.; Zhang, Z.; He, S.; Bo, X. Prediction of DTIs for High-Dimensional and Class-Imbalanced Data Based on CGAN. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, Madrid, Spain, 3–6 December 2018; pp. 788–791. [CrossRef]

24. Quintana, M.; Miller, C. Towards Class-Balancing Human Comfort Datasets with GANs. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, New York, NY, USA, 13–14 November 2019; pp. 391–392. [CrossRef]
25. Xu, L.; Veeramachaneni, K. Synthesizing Tabular Data Using Generative Adversarial Networks. *arXiv* **2018**, arXiv:1811.11264.
26. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional Gan. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 7333–7343.
27. Wang, J.; Yan, X.; Liu, L.; Li, L.; Yu, Y. CTTGAN: Traffic Data Synthesizing Scheme Based on Conditional GAN. *Sensors* **2022**, *22*, 5243. [CrossRef]
28. Imran, M.; Mahmood, A.M.; Qyser, A.A.M. An Empirical Experimental Evaluation on Imbalanced Data Sets with Varied Imbalance Ratio. In Proceedings of the International Conference on Computing and Communication Technologies, Hyderabad, India, 11–13 December 2014; pp. 1–7. [CrossRef]
29. Hwang, C.H. Resolving CTGAN-Based Data Imbalance for Commercialization of Public Technology. *J. Korea Inst. Inf. Commun. Eng.* **2022**, *26*, 64–69. [CrossRef]
30. Patki, N.; Wedge, R.; Veeramachaneni, K. The Synthetic Data Vault. In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016; pp. 399–410. [CrossRef]
31. Chang, C.C.; Lin, C.J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
32. Cox, D.R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. B Stat. Methodol.* **1958**, *20*, 215–232. [CrossRef]
33. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
34. Haykin, S. *Neural Networks and Learning Machines*; Pearson Education Upper Saddle River: Hoboken, NJ, USA, 2009; Volume 3.
35. Jiang, Z.; Yang, J.; Liu, Y. Imbalanced Learning with Oversampling Based on Classification Contribution Degree. *Adv. Theory Simul.* **2021**, *4*, 2100031. [CrossRef]
36. Puri, A.; Gupta, M.K. Comparative Analysis of Resampling Techniques under Noisy Imbalanced Datasets. In Proceedings of the 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 27–28 September 2019; Volume 1, pp. 1–5. [CrossRef]
37. Nugraha, R.A.; Pardede, H.F.; Subekti, A. Oversampling Based on Generative Adversarial Networks to Overcome Imbalance Data in Predicting Fraud Insurance Claim. *Kuwait J. Sci.* **2022**, *49*, 1–12. [CrossRef]
38. García-Vicente, C.; Chushig-Muzo, D.; Mora-Jiménez, I.; Fabelo, H.; Gram, I.T.; Løchen, M.-L.; Granja, C.; Soguero-Ruiz, C. Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors. *Appl. Sci.* **2023**, *13*, 4119. [CrossRef]
39. Douzas, G.; Bacao, F. Effective Data Generation for Imbalanced Learning Using Conditional Generative Adversarial Networks. *Expert Syst. Appl.* **2018**, *91*, 464–471. [CrossRef]
40. Ahsan, R.; Shi, W.; Ma, X.; Lee Croft, W. A Comparative Analysis of CGAN-Based Oversampling for Anomaly Detection. *IET Cyber-Phys. Syst. Theory Appl.* **2022**, *7*, 40–50. [CrossRef]
41. Son, M.; Jung, S.; Jung, S.; Hwang, E. BCGAN: A CGAN-Based Over-Sampling Model Using the Boundary Class for Data Balancing. *J. Supercomput.* **2021**, *77*, 10463–10487. [CrossRef]
42. Shelke, M.S.; Deshmukh, P.R.; Shandilya, V.K. A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *Int. J. Recent Trends Eng. Res.* **2017**, *3*, 444–449. [CrossRef]
43. Zhang, T.; Chen, J.; Li, F.; Zhang, K.; Lv, H.; He, S.; Xu, E. Intelligent Fault Diagnosis of Machines with Small & Imbalanced Data: A State-of-the-Art Review and Possible Extensions. *ISA Trans.* **2022**, *119*, 152–171. [CrossRef] [PubMed]
44. Sharma, A.; Singh, P.K.; Chandra, R. SMOTified-GAN for Class Imbalanced Pattern Classification Problems. *IEEE Access* **2022**, *10*, 30655–30665. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.